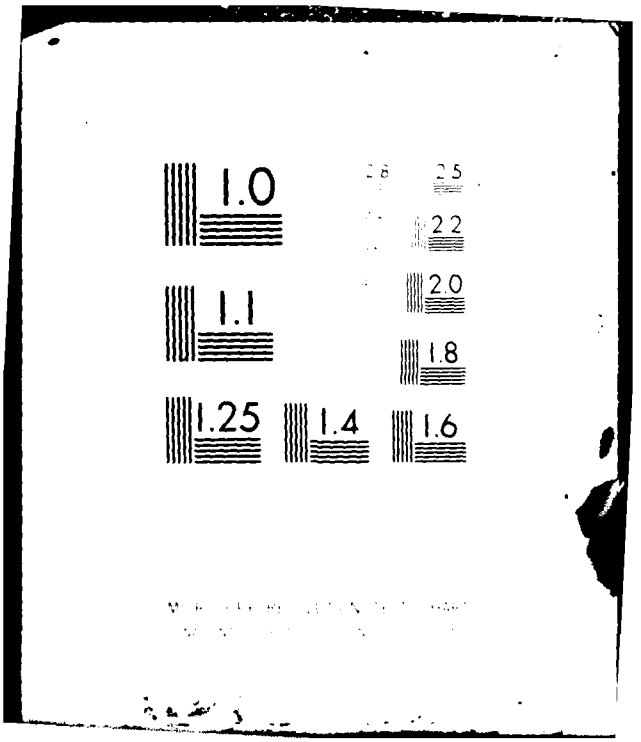
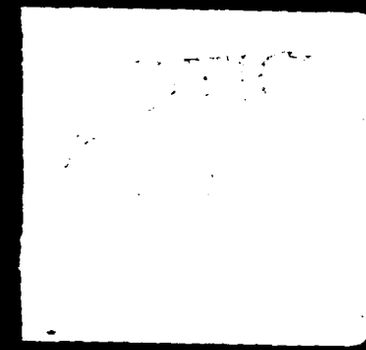
AD-A115 940 CHICAGO UNIV IL CENTER FOR DECISION RESEARCH F/G 5/10
A THEORY OF DIAGNOSTIC INFERENCE. I. IMAGINATION AND THE PSYCHO--ETC(U)
JUN 82 H J EINHORN, R M HOGARTH N00014-81-K-0314
UNCLASSIFIED 2 NL


1.0
1.1
1.25
1.4
1.6
2.0
1.8
2.2
2.5

AD A115940

CENTER FOR DECISION RESEARCH



Graduate School of Business
The University of Chicago

(12)

A THEORY OF DIAGNOSTIC INFERENCE: I. IMAGINATION AND THE PSYCHOPHYSICS OF EVIDENCE

Hillel J. Einhorn and Robin M. Hogarth
University of Chicago
Graduate School of Business
Center for Decision Research

June 1982

DTIC ELECTE JUN 22 1982 H

DISTRIBUTION STATEMENT A
Approved for public release;
Distribution Unlimited

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE	READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER: 2	2. GOVT ACCESSION NO.: AD A115440	3. RECIPIENT'S CATALOG NUMBER:
4. TITLE (and Subtitle): A THEORY OF DIAGNOSTIC INFERENCE: I. IMAGINATION AND THE PSYCHOPHYSICS OF EVIDENCE.	5. TYPE OF REPORT & PERIOD COVERED: Technical report
	6. PERFORMING ORG. REPORT NUMBER:
7. AUTHOR(s): Hillel J. Einhorn and Robin M. Hogarth	8. CONTRACT OR GRANT NUMBER(s): N00014-81-K0314
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Center for Decision Research, University of Chicago, 1101 East 58th Street Chicago, Illinois 60637	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: NR 197-071
11. CONTROLLING OFFICE NAME AND ADDRESS: Office of Naval Research 800 North Quincy Street Arlington, VA 22217	12. REPORT DATE: June 1982
	13. NUMBER OF PAGES: 72
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)	15. SECURITY CLASS. (of this report): Unclassified
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Diagnosis; decision making; evidence; attention; anchoring and adjustment.

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Diagnostic inference involves the assessment and generation of causal hypotheses to account for observed outcomes/evidence. The importance of diagnosis for prediction, defining "relevant" variables, and illuminating the nature of conflicting metaphors in inference is first discussed. Since many diagnostic situations involve conflicting evidence, a model is developed for describing how people assess the likelihood that one of two hypotheses is true on the basis of varying amounts of evidence for each. A

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE S/N 0102-LF-014-6601

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

20.

central notion is that one compares the evidence at hand with the evidence that "might have been." This is modeled via an anchoring and adjusting process where the anchor represents "what is" and the adjustment is based on imagining a contrast case for comparison. Four aspects of this model are then considered. The relation between evidentiary strength and amount of evidence (the evidence function) is shown to mimic a set of power functions. Moreover, the form of the function implies that people will trade-off relative frequency (p) for amount of evidence (n) at small n; that the absolute amount of evidence affects evidentiary strength independent of p; and that "over" and "underweighting" of probabilities decreases as amount of evidence increases. Furthermore, the model specifies when attentional shifts due to rephrasing likelihood questions will lead to the subadditivity of complementary probabilities (focus effect). Experimental data are presented that conform closely to the model's predictions. Extensions of the model are then developed to consider "diffusion" and "missing evidence" effects. The former involves changes in evidentiary strength that are caused by varying the number and specificity of alternative hypotheses. The latter involves the diagnostic impact of missing evidence presumed to have causal significance. Experimental evidence is also presented concerning both effects. The implications of the work are then discussed with respect to: (1) the role of imagination in inference; (2) the prevalence and importance of anchoring and adjustment strategies in judgment; and (3) the importance of attention in evaluating evidence. Finally, normative implications of the theory are discussed.

Accession For
NTIS GRA&I
DTIC TAB
Unannounced
Justification
By
Distribution/
Availability Codes
Dist
A

DTIC
COPY
INSPECTED
2

A Theory of Diagnostic Inference: I. Imagination and the Psychophysics of Evidence

Hillel J. Einhorn and Robin M. Hogarth
University of Chicago
Graduate School of Business
Center for Decision Research

The psychological study of inference in decision making has been dominated by the metaphor of "man as an intuitive statistician" (cf. Peterson & Beach, 1967). However, other metaphors that contrast strongly with the language and concepts of statistics could also be used; for example, one could consider man as an intuitive historian, detective, lawyer/judge, or clinician. These alternative metaphors differ from the statistical model in at least four important ways: (1) Role of causality - the statistical model does not formally consider causal ideas in its language. Indeed, concepts such as randomness, urn models, sampling, and the like, highlight the acausal nature of statistical inference. Moreover, statisticians do not encourage causal thinking, as for example in warning that correlation does not imply causation (although what does imply causation is never made clear). Contrast this view with that of the historian, detective, lawyer, or clinician. These activities are intimately concerned with causality since they involve interpreting and making sense of outcomes via examination of the causal processes that produced them; (2) Direction of inference - statistical methods are greatly concerned with forecasting events or consequences and can thus be characterized as involving "forward" inference. On the other hand, the historian, detective, etc., are generally engaged in "backward" inference; that is, one must go backwards in time to construct explanations that make outcomes seem inevitable. We call inferences that are both backward and causal, "diagnostic," and use the term "diagnosis" as a convenient summary term for comparison with the statistical metaphor; (3) Domain of inference -

statistical concepts such as averages, variability, relative frequency, population, and so on, clearly indicate that the domain to which inferences are being made are aggregates of some sort. Therefore, one is concerned with the general case or with classes of cases. Indeed, the importance of set theory for defining statistical concepts emphasizes that individual cases are to be considered as members of sets or subsets of like members. Moreover, when this cannot be easily accomplished, as when considering one-of-a-kind events (such as the likelihood of a Russian invasion of Poland), controversy exists regarding the meaning and meaningfulness of probability statements. Now consider the domain of inference of the lawyer, detective, or historian. Here one is concerned solely with the specific case--did Mr. X commit the crime? Is Mr. Y responsible for the accident? What were the causes of World War I? While appeal can be made to general laws or principles in answering such questions, the relevance of such evidence is often questioned on the grounds that the specific case is not a member of this or that class. Indeed, it is not uncommon to hear that certain events (and people) are considered, "in a class by themselves;" (4) Role of quantitative rules for combining evidence - statistical theory relies on quantitative rules for evaluating information. Of particular importance is the use of Bayes' theorem in the updating of one's prior beliefs on the basis of new evidence. In fact, the development of Bayesian theory and methods has been quite substantial in recent years, attesting to its importance as a tool for evaluating and combining evidence. On the other hand, the lack of quantitative rules for doing causal-backward inference in specific cases is striking. Moreover, when one considers that diagnosis is a highly constructive process, the difficulty of formalizing it becomes apparent. That is, by "constructive" we mean that the diagnostician must create plausible causal scenarios that are

sufficient for explaining outcomes that have occurred. Furthermore, in doing this, scenarios or explanations should "make the data as redundant as possible" (Bruner, 1957). An important way to accomplish this is through the continued revision, expansion, and synthesis of hypotheses and models to make the data that has happened appear most likely to have happened. How this is achieved is poorly understood, yet it seems safe to say that it does not rest on the use of quantitative, formal rules. In fact, this aspect of the diagnostic process has received little attention in statistical models concerned with the combining of evidence (cf. Gettys and Fisher, 1979). For example, consider Agatha Christie's Murder on the Orient Express. Inspector Poirot starts with what seems to be 12 mutually exclusive and exhaustive suspects regarding a murder that has occurred on board the train. But, as we discover to our surprise, he correctly figures out that all 12 did it. While it could be argued that Poirot considered all the various combinations of suspects as alternative hypotheses (i.e., $\sum_{i=1}^{12} {}_{12}C_i$), this is clearly stretching information processing requirements, even for the remarkable Poirot. Similarly, consider a physician who holds several hypotheses regarding what disease a patient has and arrives at a diagnosis in which the patient has disease A with complications due to disease B. Since an a priori enumeration of all possible combinations of diseases is impossible, changing or synthesizing hypotheses rather than changing the probability of fixed hypotheses must be involved.

We now discuss several implications of considering diagnosis as a constructive process that is causal, backward in its direction, and concerned with specific cases.

(a) Diagnosis precedes prognosis. Imagine you have been asked to evaluate the research output of a younger colleague being considered for promotion. Your colleague has produced 11 papers; of these the first 9 (in

chronological order) represent competent, albeit unexciting scholarly work. On the other hand, the latter 2 papers are quite different; they are innovative and suggest a creativity and depth of thought absent from the earlier work. What should you do? As someone who is aware of regression fallacies, you might consider the two outstanding papers as outliers from a stable generating process and thus predict regression to the mean. Alternatively, you might consider the outstanding papers as "extreme" responses that signal (or are diagnostic of) a change in the generating process. If this were the case, you should be predicting future papers of high quality. If one asks what is the nature of the signaling in this case, it is obvious that the chronological order of the papers is crucial for making the diagnosis. Indeed, imagine that the outstanding papers were the first two that were written; or consider that they were the second and sixth. Each of these cases suggests a different diagnosis and perhaps a different prognosis.

The above example illustrates that prognoses depend crucially on prior diagnoses, even when such diagnoses are carried out quickly and with little awareness. Moreover, little guidance can be offered for helping one to decide what is the appropriate model that has generated, and presumably will continue to generate, the variable of interest. In fact, the validation of diagnoses is fraught with difficulties since outcomes used to assess the accuracy of predictions may be "contaminated" by actions taken on the basis of the predictions (see e.g., Einhorn & Hogarth, 1978). In any event, the interplay between forward and backward inference is important and can be eloquently summarized by Kierkegaard's remark that, "Life can only be understood backwards; but it must be lived forwards."

(b) Conflicting metaphors. When statistical methods are applied in various content areas, there are often disputes about its applicability. Such

disputes seem to rest on a lack of awareness that a particular metaphor may not be equally applicable in a variety of contexts. In particular, because statistical inferences are acausal, forward, formal, and deal with aggregates, they are most likely to conflict with areas in which inferences are causal, backward, informal, and deal with specific cases. Such an area is the legal field and indeed, much controversy exists over the use of probabilistic ideas in the judicial system. For example, consider the use of probabilistic evidence in the following scenario: 1000 people attend a concert where only one ticket has been sold. Thus, 999 people have broken the law by illegally gaining admittance to the concert hall. A security guard randomly chooses one person to arrest and that person is then brought to trial. The prosecution argues that the probability of guilt of the accused is .999, which is beyond a reasonable doubt. Should the person be found guilty on the evidence presented? If not, why not?

(c) Relevant variables and information. The question just posed raises the issue of defining "relevant" information. That is, the concept of "equity," the lack of specific-causal evidence regarding the arrested person, and so on, are relevant when the problem is put in a legal setting although they may be irrelevant in other contexts. Therefore, relevance can only be understood in relation to some model or metaphor that delimits what is attended to (the figure) from what is not (the ground). From a psychological perspective, the speed with which diagnosis is performed and the frequent lack of awareness that a delimiting process has occurred, is of particular importance. Indeed, the use of metaphor illustrates just this point. For example, a model of the brain as computer has frequently been used to emphasize processes of informational input, retrieval, and computation. However, one could also consider the brain as a muscle or a sponge. Such images

immediately bring to mind quite different processes and variables (e.g., weakening and strengthening with use, the strain of thinking, soaking up information, etc.). However, note that information that is relevant in one model is not of equal importance or relevance in others. Moreover, it is only by considering multiple models that one becomes aware of this. Indeed, this has been precisely our purpose in considering the various metaphors used to describe inference in decision making.

Plan of the paper

While it is obvious that evidence is central to inference, the meaning of evidence must be considered more broadly than is typically the case in current formal models. In particular, evidence is more than knowledge of outcomes since outcomes are diagnostic of the processes that produced them. Therefore, an important aspect of evidence is the degree to which it stimulates imagination so that new or different interpretations are constructed. Indeed, diagnosis differs from simple categorization in that the former is constructive while the latter is not. These ideas are discussed within the context of a model for evaluating the strength of conflicting evidence. The model deals with assessing the likelihood that one of two positions is true on the basis of varying amounts of evidence for each. Central to the model is the idea that one compares the evidence that one has with the evidence that might have been. Thus, one's ability to use imagination for constructing "what might have been" is an essential component in assessing evidentiary strength.

We first present the formal model in the next section and consider the relation between evidentiary strength and amount of evidence on hand (called "the evidence function"). We then consider how attentional shifts can lead to the subadditivity of complementary probabilities (focus effects). Experi-

mental evidence is then presented regarding both the evidence function and focus effects. Thereafter, we consider "diffusion" and "missing evidence" effects. The former involves changes in the strength of evidence caused by varying the number and specificity of alternative hypotheses. The latter involves the diagnostic impact of missing evidence presumed to have causal significance. Experimental evidence regarding both effects is also presented. Finally, we discuss the general implications of the above with respect to: (1) The role of imagination in inference; (2) the prevalence and importance of anchoring and adjusting strategies in judgment; (3) the importance of attention in evaluating evidence; and (4) the normative implications of our theoretical position.

A Model for the Evaluation of Evidence

We begin by assuming that the evaluation of evidence generally involves some implicit or explicit background against which evidentiary strength is assessed. As such, the evaluation of evidence is similar to judgment in other domains where the judged object is made more or less salient by changes in the background against which it is perceived. Consider, for example, the relation between figure and ground in perception (see e.g., Allport, 1955), the effect of decision frames in choice (Tversky & Kahneman, 1981), or the ability to distinguish signals from noise (Green & Swets, 1966). Furthermore, a striking feature of such phenomena is that in judging the clarity of percepts, people are only aware of the net effect of two forces, i.e., the clarity of the figure and the degree of diffuseness of the ground. Indeed, the former is judged by reference to the latter. By analogy with these sensory phenomena, we argue that the evaluation of evidence will result from the net effect of arguments for and against a particular proposition. Arguments for the

proposition will be the focus of attention and thus represent figure against the ground of counter-arguments. However, the manner in which the counter-arguments are structured will affect the net strength of the evidence.

Consider that there are n equally strong pieces of evidence that consist of f favorable and c unfavorable or "con" arguments. Let $S_n(f:c)$ be defined as the net strength of f "for" and c "con" pieces of evidence where the total amount of evidence at hand is $n = f+c$. Moreover, let $p = f/n$, which is simply the proportion of favorable evidence to the total amount on hand. For example, imagine that there has been a hit-and-run accident where f witnesses say the offending car was blue while c witnesses claim it was green. We are interested in the evaluation of the likelihood that a blue car caused the accident as a function of f, c, p, and n. Note that the situations we are considering involve inferences that are causal, backward, and concerned with a specific case. Therefore, our model of how these inferences are made will not be identical to standard probability approaches.

The model we propose is based on an anchoring and adjustment process (Tversky & Kahneman, 1974); specifically, we assume that in evaluating conflicting evidence, one first anchors on p, the proportion of favorable to total evidence on hand. However, adjustments to p will be made on the basis of four factors: (1) the amount of total evidence at hand, n; (2) whether one is evaluating a particular hypothesis or its complement; (3) the number and specificity of alternative hypotheses; (4) perceptions of missing evidence. We now consider each of these in turn.

(1) Effects of amount of evidence on hand (n): The evidence function

We propose that there is a tradeoff between p and n such that one would accept less p for greater n. For example, consider that there were only two witnesses who both said that the hit-and-run car was blue vs. a situation in which 9 witnesses said blue and one said green. Many people find the latter evidence stronger than the former in supporting the proposition that a blue car caused the accident. Why? We argue that when the total amount of evidence is meager, it is quite easy to imagine a different result by simply changing one piece of evidence. Thus, an outcome of (2:0) could easily be (1:1); or, (2:1) becomes (1:2) if only one witness changes his/her mind. Therefore, we propose that people will anchor on p and then adjust for n by imagining a worse case in which one f is shifted to one c.[1] The model can be formally stated as follows. Let,

$$S_n(f:c) = a_1\left(\frac{f}{n}\right) + a_2\left(\frac{f-1}{n}\right) \qquad (1)$$

where, a_1 = weight for the anchor
a_2 = weight for the adjustment

Moreover, we assume that, $a_1 + a_2 = 1$, which makes S_n a weighted average of p and the imagined worse case. The rationale for this has been discussed by Lopes (Note 1), who notes that averaging models in judgment may reflect underlying anchoring and adjustment strategies. Therefore, we expect that S_n will fall somewhere between p and (f-1)/n, depending on the relative weighting of the anchor and adjustment components. Equation (1) can now be rewritten as,

$$S_n(f:c) = \frac{a_1 f + a_2 f - a_2}{n} = \frac{f(a_1 + a_2) - a_2}{n} = \frac{f - a_2}{n} = p - \frac{a_2}{n} \qquad (2)$$

Equation (2) highlights the nature of the adjustment process due to n; when $a_2 = 0$, n plays no role and evidence is solely a function of p. However, when $a_2 > 0$, adjustment is largest when n is small. For example, consider an outcome of (2:1) with $a_2 = .4$. The net strength of such evidence is $.67-(.4)/3 = .54$. However, as n increases, the adjustment decreases so that in the limit, $S_n \to p$ as $n \to \infty$. The model therefore predicts that as the total amount of evidence increases, net strength approaches p as an asymptote, with a rate determined by a_2. However, when $p = 0$, equation (2) results in net strength being negative, which is conceptually undesirable. Accordingly, the anchoring and adjustment process outlined above is modified when p is at, or close to zero (i.e., $0 \leq p \leq p_c$, where p_c is some small value of p). Specifically, instead of imagining a worse case in the adjustment process, we hypothesize that an imagined better case provides the contrast for the adjustment. Furthermore, the nature of the adjustment is the reverse of the process represented in (2); i.e., one c is moved to one f. This results in $S_n'(f:c)$, where,

$$S_n'(f:c) = a_1 \frac{f}{n} + a_2 \left(\frac{f+1}{n}\right) \qquad (3)$$
$$= p + \frac{a_2}{n}$$

Note that (3) is the same as (2) except that the sign of the a_2/n term is positive rather than negative. Furthermore, when $p = 0$, $S_n'(0:c) = a_2/n$, which has the property that as $n \to \infty$, net strength approaches an asymptote of zero. Indeed, for all $p \leq p_c$, $S_n' \to p$ as $n \to \infty$. While (2) and (3) both predict that net strength will asymptote at p, the asymptotes are approached from below in one case (equation (2)), and from above in the other (equation

(3)). The implications of this will be discussed in the next section. Finally, we can express the general relation between S_n and n, hereafter called the "evidence function," as,

$$S_n(f{:}c) = p + \beta\left(\frac{a_2}{n}\right) \qquad (4)$$

where, $\beta = \begin{cases} 1 \text{ if } p \leq p_c \\ -1 \text{ if } p > p_c \end{cases}$

Figure 1 shows the functional relation between $S_n(f{:}c)$ and n for varying levels of p (using $a_2 = .3$ for illustration). There are several interesting aspects to note: (a) The relations between S_n and n mimic a set of power functions. As pointed out by Latane (1981), the difficulty

Insert Figure 1 about here

of combining mixtures of power functions for obtaining a psychophysics of conflicting arguments has not yet been overcome. However, the rather simple formulation of equations (2) and (3) does manage to yield power-like functions that deal directly with conflicting evidence; (b) The trade-off between n and p can be determined by considering the points that lie on a line at which S_n is a constant (or nearly so). For example, compare (2:0), with a net strength of .85 ($a_2 = .3$), to (9:1), with $S_n = .87$. While the two strengths are not identical, their closeness should lead to judgments of comparable impact. As Figure 1 makes clear, n and p will only trade-off within certain ranges of p at low levels of n. (c) Trade-offs between n and p will be further affected by the fact that when evidence is discrete, certain levels of p and n cannot exist; e.g., $p = .9$

when $n < 10$. The parts of each function relating S_n to n that cannot represent discrete pieces of evidence are shown by dashed lines.[2] The implication of this is that trade-offs between p and n may occur by moving to the nearest part of the function for which a discrete amount of evidence exists. For example, evidence of (2:0) has the same net strength as (5.4:.6), which is characterized by $p = .9$, $n = 6$. However, since these latter values are impossible in the discrete case, one might make a trade-off at 9:1 although $S_n(9:1) > S_n(2:0)$.

We now consider three important implications of our model: (i) The model expressed in (2) seems to capture the distinction made by Keynes (1921) between the probability of a hypothesis (p) and what he called the "weight of evidence." He stated,

> The magnitude of the probability. . .depends upon a balance between what may be termed the favourable and the unfavourable evidence; a new piece of evidence which leaves this balance unchanged, also leaves the probability of the argument unchanged. But it seems that there may be another respect in which some kind of quantitative comparison between arguments is possible. This comparison turns upon a balance, not between the favourable and unfavourable evidence, but between the absolute amounts of relevant knowledge and of relevant ignorance respectively. (Keynes, 1921, p. 71, original emphasis).

Keynes went on to say that while additional relevant evidence can lower or increase the probability of a hypothesis it always increases its "weight" since, "we have a more substantial basis upon which to rest our conclusions." The distinction between evidentiary weight and probability is precisely captured in (2) since S_n increases with the absolute amount of evidence (n) even though p is held constant. Thus, Keynes' distinction can be given a quantitative representation in our model; (ii) There has been much interest in the relation between measures of subjective uncertainty and probabilities based on relative frequency (e.g., Estes, 1976). It is therefore useful to

examine what our model predicts to be the functional relation between S_n and p. This is shown in Figure 2 (assuming that $a_2 > 0$). Since amount of

Insert Figure 2 about here

evidence plays a crucial role in our model, the functions relating S_n and p are shown when n is large, moderate, and small.[3] Note that in general, low probabilities are "overweighted" (i.e., $S_n > p$ when $p < p_c$) whereas probabilities above p_c are "underweighted" (i.e., $S_n < p$). A similar effect has been hypothesized by Kahneman and Tversky (1979). In discussing how uncertainty impacts on choice, they define what they call "decision weights," which are related to explicit probabilities such that small probabilities are overweighted and all others are underweighted. Although their function is undefined at the end points of $p = 0$ and 1, and accepting the fact that the contexts of the two theories are different, the similarity of both treatments of uncertainty is striking. On the other hand, in our model, as n increases, over-and underweighting of p decreases. Indeed, when n is large, $S_n \approx p$, thereby leading to a standard probability model of evidence as a special case; (iii) Because p is a central part of the basic model, our formulation implies the following: For a majority or neutral position (i.e., $2f \geq n$), the addition of positive evidence has less effect on the net strength of evidence than the reduction of an equal amount of negative evidence. For example, compare the addition of one positive argument to make evidence of (3:2) into (4:2), vs. the reduction of one negative argument to yield (3:1). According to probability theory and our model, (3:1) is stronger evidence than (4:2). However, the two models diverge when the deletion of negative arguments results in a loss of n such that large downward adjustments to p result. For example, consider initial evidence of (1:1) and compare (2:1) to (1:0). If $a_2 = .4$, $S_n(2:1) = .58$ and $S_n(1:0) =$

.60, which are much closer than would be the case if probability were used as a measure of evidentiary strength. Therefore, while the deletion of negative evidence generally strengthens a majority or neutral position more than the addition of positive evidence, deletions that substantially reduce n can lower net strength and thus work against increases in p. Now consider a minority position (i.e., $2f < n$). In this case, the addition of positive evidence is superior to the reduction of the same amount of negative evidence. For example, the position (2:4) becomes either (3:4) or (2:3) with the addition and deletion of 1f and 1c, respectively. However, our model again departs from probability theory when n is small and $a_2 > 0$. Consider evidence of (3:4) and its change to (4:4) vs. (3:3). While a probability approach would treat these as equal, our model evaluates $S_n(4:4) > S_n(3:3)$ (for $a_2 > 0$). Thus, the effectiveness of deleting negative evidence is again attenuated by reducing n.

(2) The focus effect

Imagine that you were asked to evaluate the likelihood that some hypothesis is true on the basis of f positive pieces of evidence and c negative pieces ($f+c = n$ being the only evidence available). Now consider that the question was reversed in the following way: how likely is it that the complementary hypothesis is true on the basis of c pieces of evidence for it, and f pieces in favor of the first hypothesis. Note that the evidence is identical in both questions and standard probability approaches would treat the two probabilities as complements. That is, the probability of the first hypothesis being true is f/n, the probability of the complementary hypothesis being true is c/n, and $(f/n) + (c/n) = 1$. However, the evaluation of evidence may be affected by how attention is directed to a

particular hypothesis via the wording of the question. In particular, we hypothesize that attentional shifts due to focusing on the evaluation of one or other hypothesis will result in a "focus effect." More formally, a focus effect is defined when,

$$S_n(c:f) < 1 - S_n(f:c) \qquad \text{or,} \qquad (5a)$$

$$S_n'(c:f) < 1 - S_n(f:c) \qquad (5b)$$

For example, consider that evidence in a trial consists of 4 pieces that support guilt (f) and one that supports innocence (c). The question, "How likely is the person guilty on the basis of (4:1)?" will be evaluated according to (2) as $S_n(4:1) = .8 - \frac{a_2}{5}$. However, the question, "How likely is the person innocent on the basis of (1:4)?" will not generally be $1 - S_n(4:1)$. The reason for this follows directly from the form of the evidence function and the assumed anchoring and adjustment process. That is, in answering the question about guilt, one anchors on p and adjusts downward for n; similarly, in answering the question about innocence, one anchors on $(1-p)$ and also adjusts downward for n. The two downward adjustments lead to the inequality shown in (5a).

We now formally consider the conditions that lead to focus effects. In order to do so, first note that,

$$S_n(c:f) = (1-p) - \frac{a_2}{n} \qquad \text{and,} \qquad (6a)$$

$$S_n'(c:f) = (1-p) + \frac{a_2}{n} \qquad (6b)$$

We distinguish situations where both p and $(1-p)$ are greater than p_c from

those where either is less than or equal to p_c. Thus, in the first case

$$S_n(f:c) + S_n(c:f) = \left\{p - \frac{a_2}{n}\right\} + \left\{(1-p) - \frac{a_2}{n}\right\}$$

$$= 1 - \frac{2a_2}{n} \qquad (7)$$

such that the focus effect only occurs if $a_2 > 0$.

However, in the second case, e.g., $(1-p) \leq p_c$, then

$$S_n(f:c) + S'_n(c:f) = \left\{p - \frac{a_2}{n}\right\} + \left\{(1-p) + \frac{a_2}{n}\right\}$$

$$= 1 \qquad (8)$$

and *no* focus effect occurs, irrespective of the value of a_2.

We now discuss two interesting aspects of the focus effect: (i) Complementary net strengths can be subadditive. This implication has important similarities to other treatments of uncertainty in various theories of choice (see especially Kahneman & Tversky, 1979) and inference. Of particular relevance to the present model is Cohen's work on the relations between the completeness of evidence and the complementation principle in standard probability theory (see Cohen, 1977, chapter 3). Using Keynes' distinction between the probability and weight of evidence, Cohen points out that when one considers an incomplete system, the lower benchmark that one puts on provability is not necessarily disprovability, but non-provability. For example, if one has a meager amount of circumstantial evidence supporting a particular scientific theory such that the probability of the theory's truth is .2, does that imply that the theory is false with p = .8? One might rather say that the theory is not proven (in a probabilistic sense) as opposed to saying that

there is a 4:1 chance it is wrong. As Cohen states,

> . . .in everyday life we very often have to form beliefs about individual matters of fact in a context of incomplete information. We need something better then than a concept of probability which is conclusively applicable to our inferences only on the assumption that we already have all the relevant premises. (1977, p. 39)

Furthermore, the idea that the complement of statements can lead to "not-proved" rather than "disproved," seems to be deeply imbedded in the Anglo-American legal system. Indeed, in Scottish law, defendants are either found guilty, not-guilty, or "not proven." The last category is reserved for those cases where the evidence is too meager to allow for a judgment of guilt or innocence. Now consider how our model deals with the issue of complementarity and the completeness of evidence. When p and $(1-p)$ are both greater than p_c, it follows from equation (7) that as the amount of evidence increases $(n \to \infty)$, complementation would be expected in the limit. However, if $a_2 = 0$ (there being no adjustment for n), complementation holds even at small amounts of evidence. On the other hand, when p or $(1-p)$ is equal to or less than p_c, complementation holds at all levels of n--see equation (8); (ii) The focus effect discussed here resembles several effects considered in Tversky's model of similarity judgment (1977). For example, the judged similarity of objects a and b may not be the same as the similarity of b and a, as in, "a man is like a tree" vs. "a tree is like a man." This asymmetry occurs because attention is focused on one object as subject and the other as referent. When subject and referent are reversed, attention is shifted and asymmetric similarity judgments occur. Furthermore, when people judge the similarity of objects, they focus attention on the objects' common elements while judgments of dissimilarity focus attention on unique elements. This leads to the interesting case in which two objects are judged as highly similar and highly dissimilar, depending on the question asked. The focus

effect discussed here involves attentional shifts in a similar way; namely, one's attention is first directed to p as the anchor in one form of the question while the anchor in the other question is $(1 - p)$. The subsequent downward adjustments for n lead to the focus effect.

We now discuss several experimental tests of the above model by empirically examining the form of the evidence function and the extent to which focus effects occur in the data.

Method: Experiment 1

Subjects

Thirty-two subjects were recruited through an ad in the Univesity newspaper which offered $5 an hour for participation in an experiment on judgment. The median age of the subjects was 24, their educational level was high (mean of 4.4 years of formal post-high school education), and there were 16 males and 16 females.

Stimuli

The stimuli consisted of a set of scenarios that involved a hit-and-run accident seen by varying numbers of witnesses. Moreover, of the n witnesses to the accident, f claimed that it was a green car while c claimed it was a blue car. A typical scenario was phrased as follows:

> "An automobile accident occurred at a street corner in downtown Chicago. The car that caused the accident did not stop but sped away from the scene. Of the n witnesses to the accident, f reported that the color of the offending car was green, whereas c reported it was blue. On the basis of this evidence, how likely is it that the car was green?"

Each scenario was printed on a separate page and contained a 0-100 point rating scale that was used by the subject to judge how likely the accident was caused by a particular colored car. Each stimulus contained the same basic

story but varied in the total number of witnesses (n), the number saying it was a green (f) or a blue car (c), and whether one was to judge the likelihood that the majority or minority position was true. In order to sample a wide range of values of n and p, 40 combinations were chosen as follows: for p = 1, n = 2, 6, 12, 20; p = .89, n = 9, 18, 27; p = .80, n = 5, 10, 15, 20, 25; p = .75, n = 4; p = .67, n = 3, 6, 9, 12, 15, 18, 24; p = .60, n = 5, 10; p = .50, n = 2, 8, 12, 20; p = .40, n = 5, 10; p = .33, n = 6, 9, 18; p = .25, n = 4; p = .20, n = 5, 10; p = .11, n = 9, 18; p = 0, n = 2, 6, 12, 20. In addition, 8 stimuli were given twice to ascertain test-retest reliability. Thus, the total number of stimuli was 48, and they were arranged in booklet form.

Procedure

When the subjects entered the laboratory, they were told that the experiment involved making inferential judgments. Furthermore, it was stated that if they did well in the experiment (without specifying what this meant), it was likely that they would be called for further experiments. Given the relatively high hourly wage that was paid, this was thought to increase the incentive to take the task seriously. In order to avoid boredom and to reduce the transparency that judgments of complementary events were sometimes called for, subjects were given 4 sets of 12 stimuli and, after completing each set, they performed a different task. All stimuli were randomly ordered within the four sets. Subjects could take as much time as they needed and they were free to make as many (or as few) calculations as they wished. After completing the task, all subjects filled out a questionnaire regarding various demographic variables.

Results

Since the study is a full within-subjects design (each subject rated 48 stimuli), data are available to test the model on each subject. However, by averaging over subjects, one can also test the model on the aggregate data. We first discuss the results for the aggregate data and then consider the individual analyses.

Aggregate analyses

Recall that for each subject, 8 stimuli were given twice so that test-retest reliability could be assessed. This was done in two ways: (1) the correlation between judgments of the same stimuli, within each subject (N = 8), was computed. The mean of these correlations was .93, with 26 of the 32 coefficients greater than .90; (2) each subject was considered a replication with 8 responses and the correlation between judgments for identical stimuli, over subjects (N = 256 = 32 subjects x 8 responses), was .91. Clearly, the reliability of the judgments was high, regardless of the computational method.

In order to test our model on the aggregate data, the 32 responses for each of the 48 stimuli were averaged to obtain what we call, "mean net strength," $\bar{s}_n(f:c)$. Using $\bar{s}_n(f:c)$ as our new dependent variable, the two parameters in the model, a_2 and p_c have to be estimated from the data. This makes it necessary to incorporate random error explicitly in our formulation. Accordingly, we used a regression approach where $\bar{s}_n(f:c)$ is assumed to consist of a predicted part, $\hat{\bar{s}}_n(f:c)$, and a random error component, ε; i.e.,

$$\bar{s}_n(f:c) = \hat{\bar{s}}_n(f:c) + \varepsilon \quad , \qquad (8)$$
$$\varepsilon \sim N(0, \sigma_\varepsilon^2)$$

where,

The predicted portion of $\bar{S}_n$ can be written as a regression of the form,

$$\hat{\bar{S}}_n(f{:}c) = p - \hat{a}_2\left(\frac{1}{n}\right) \quad , \quad \text{where,} \tag{9}$$

$\hat{a}_2$ = estimated weight for the hypothesized adjustment process.

However, there are two difficulties in estimating (9): (a) Recall that when $p < p_c$, the sign of a_2 is positive. To deal with this, we examined the aggregate data and found where $\bar{S}_n > p$ at small n. This point was defined as p_c and a_2 was assumed to be positive for $p < p_c$. For our data, $p_c = .20$; (b) A statistical problem in estimating a_2 from (9) is that p and $\frac{1}{n}$ must be highly correlated since $p = f\left(\frac{1}{n}\right)$. This makes the determination and testing of a_2 problematic. In order to handle this, the following two-step procedure was used: $\bar{S}_n$ was first regressed onto p to test for the importance of p as the anchor in the assumed evaluation process. As expected, the correlation was high ($r = .98$). However, the real test of the model concerns the adjustment process; i.e., is a_2 negative in sign and statistically significant? This led to the second step. By substituting (9) into (8) and re-arranging terms, the regression formulation can be written:

$$p - \bar{S}_n(f{:}c) = \hat{a}_2\left(\frac{1}{n}\right) + \varepsilon \tag{10}$$

By regressing the difference, $(p - \bar{S}_n)$, onto $\frac{1}{n}$, we can test whether the hypothesized adjustment process predicts the <u>differences</u> between mean net strength and p.[4] This regression was performed by constraining the

regression line through the origin (since there is no intercept in equation (10)) with the following results: the squared correlation between $(p - \bar{S}_n)$ and $1/n$ was .50 and the estimate of $\hat{a}_2 = .26$ ($F = 153.6$, $p < .0001$). In order to see just how well the model fits the data, Table 1 shows the actual and predicted mean net strengths for the 48 stimuli and Figure 3 presents a visual display of the same data.

Insert Table 1 and Figure 3 about here

In examining Table 1, note how the net strength model correctly captures the adjustment for small n, in contrast to a model based only on p. For example, when $p = 1$, $\bar{S}_n < p$, with the largest discrepancies occurring when n is small. That $\bar{S}_n$ approaches its asymptote of p with increases in n can be seen most clearly when $p = 1$, .67, .60, .50, and .40. Furthermore, when $p = 0$, S_n decreases with larger n (as predicted) and a_2/n provides a good fit to the data.

Our second set of results concerns the focus effect. The data are presented in Table 2. As predicted, when $p \leq p_c$, there are no focus effects.

Insert Table 2 about here

However, when $p > p_c$, focus effects exist and the net strength model predicts their magnitude reasonably well.

Individual analyses

The individual analyses were based on exactly the same method used in the aggregate analysis. That is, for each subject, p_c was first assessed, and then $S_n(f{:}c)$ was regressed onto p to test for the significance of the anchor in the model. The individual correlations were high and statistically

significant (correlations ranged from .89 to .99). Thereafter, $p - S_n(f:c)$ was regressed onto $1/n$ in order to estimate the a_2 parameter for each subject. These results are shown in Table 3.

Insert Table 3 about here

Note that for 22 of the 30 subjects (2 subjects were dropped because their data were too erratic), the $\hat{a}_2$ coefficients are significant and in the predicted direction. Therefore, at the individual subject level, 73% of the subjects were better fit by our model than by a simple probability formulation. Furthermore, the average $\hat{a}_2$ over the 30 subjects is .23, which is close to the value of $a_2 (= .26)$ used in the aggregate analysis.

We now consider focus effects at the individual level. While it is too cumbersome to show the data for each subject, we present the results for three subjects with different values of $\hat{a}_2$. This should give some indication that the aggregate results are also mirrored in the individual data. Table 4 presents these results. The first subject considered in the table did not

Insert Table 4 about here

have a statistically significant $\hat{a}_2$ and one can see that there are virtually no focus effects. The regression for the second subject yielded $\hat{a}_2 = .04$, which suggests systematic but small focus effects. Indeed, the data show this pattern of results. The third subject had the largest obtained value of $\hat{a}_2$; $\hat{a}_2 = .73$ (also, $p_c = .11$). Here one can see that focus effects are substantial. In addition to the qualitative prediction of focus effects, the predictions of the magnitudes of the effects are also shown. Given the difficulty of predicting a complex effect in the presence of substantial noise in the data, we view these results as encouraging. Taken together, the individual data display the same characteristics as the aggregate data,

lending further support for the theoretical model.

Experiment 2

In order to demonstrate the stability of our findings, we replicated our first study using different scenarios. It was thought that such a replication would be important in showing that our hypothesized effects generalize over different content domains. Accordingly, we developed four new scenarios that involved: (1) A bank robbery where witnesses said the robbers spoke to each other in a foreign language (German vs. Italian); (2) an FM radio station that asked listeners to identify the composer of a recorded piece of music (Gershwin vs. Beethoven); (3) experts investigating the cause of a fire (arson vs. short-circuit); and, (4) an experiment where children had to identify words flashed on a screen (ROT vs. BED). In each scenario, there were n pieces of evidence, with f supporting one position and c supporting the other. The stimuli were identical to those used in the first experiment with respect to the levels of n, p, f, and c. Therefore, the only difference between the scenarios is their substantive content.

Subjects and procedures

Thirty-two additional subjects participated in this experiment. Eight subjects were randomly assigned to each scenario condition and they judged the likelihood that one or other position was true. There were 48 stimuli as in Experiment 1, and all other procedures were identical.

Results

We only consider the aggregate analyses since our focus is on showing that the earlier results generalize over varying content. Using mean net

strength as our dependent variable, a_2 and p_c were estimated as in Experiment 1, and the results are shown in Table 5.

Insert Table 5 about here

Note that for all scenarios, $\hat{a}_2$ is statistically significant and in the predicted direction (the t-statistic is shown in parenthesis). Although there are differences in $\hat{a}_2$ over scenarios, the values are in the general range of our first experiment ($\hat{a}_2 = .26$). Moreover, if people are using an anchoring and adjusting process, our results indicate that the relative weight for the adjustment is less than for the anchor, a result that has been suggested in other work as well (e.g., Tversky & Kahneman, 1974). Column (3) shows the values of p_c, which are different for each scenario (recall that $p_c = .20$ in Experiment 1). This result is particularly interesting since it suggests that the point at which a hypothesis changes from "probable" to "improbable" depends on the particular content of the scenario. However, further work on the reasons for shifts in p_c is needed. Columns (3) and (4) show the correlations of p with $\bar{S}_n$, and $(p - \bar{S}_n)$ with $1/n$, respectively. Finally, focus effects exist in all four scenarios, in accord with the values of a_2 and p_c. For example, in the FM-station-scenario, focus effects are considerable ($\hat{a}_2 = .44$) while they are smaller in the other scenarios. Moreover, in the scenario with $p_c = .40$, focus effects were not predicted except when $p = .50$, and the results essentially confirmed this expectation.

(3) The diffusion effect

The central theme of this paper is that evidence is evaluated by its net effect. Therefore, the strength of evidence can be directly increased by either increasing the number of positive arguments or decreasing the number of negative arguments. However, evidentiary strength can also be

affected by the structure of the negative evidence against which positive evidence is compared. To illustrate, recall our hit-and-run scenario and imagine that four witnesses reported a green car and four reported the color as blue, i.e., (4G:4B). Now consider a second situation in which four witnesses reported green, two reported blue, and two reported red; i.e., (4G:2B,2R). In this second case, is it more, less, or equally likely that a green car was responsible for the accident? We hypothesize that for many people, the strength of evidence for a green car will not be the same. We call this a diffusion effect since it results from splitting or "diffusing" the total amount of negative evidence into multiple categories or hypotheses. Note that a diffusion effect violates the evaluation of evidence in standard probability theory. That is, the probability of a hypothesis should be unaffected by the number or composition of alternative hypotheses. Thus, if the probability of some hypothesis H is p, the fact that $\bar{H}$ is made up of one or more alternatives is irrelevant to the probability of H (and therefore $\bar{H}$). However, the net strength of evidence (S_n) is sensitive to the composition of alternatives since the number of alternatives, their distribution, and the like, are themselves diagnostic of some presumed causal process.

We now discuss the psychological factors that underly diffusion effects by considering how the strength of a position can be decreased or increased by diffusing negative evidence. First, imagine that in the above example, the diffusion of negative evidence resulted in a decrease in the judged likelihood that a green car caused the accident; i.e., $S_n(4:2,2) < S_n(4:4)$. How can this be explained? We suggest that multiple categories of negative evidence (holding total c constant) are diagnostic of a causal process that produces

variable outcomes. That is, when evidence points to many alternative hypotheses, some reason for this variability is sought. A particularly simple "explanation" is that the causal process generating outcomes is highly uncertain and this leads to a reduction in the strength of all positions. For example, consider diffusing (4:4) into (4:1,1,1,1), where the negative evidence consists of 4 different car colors. What kind of process could lead to the reporting of so many different categories? One possibility is that the conditions for viewing the accident were very poor; e.g., it was foggy, twilight, and so on. However, if this diagnosis were made, the strength of the f position could also be reduced. Other examples of negative diffusion effects can often be found in evaluating scientific theories. For example, a theory that has many competitors could be seen as weaker than one that has only a single competitor. The reason for this is that evidence in support of multiple alternatives suggests that the phenomenon in question is highly uncertain and the likelihood of any single position being correct is reduced.

Now imagine that the diffusing of negative evidence resulted in an increase in the strength of the f position. We call this a positive diffusion effect and offer the following rationale: When multiple categories or alternative hypotheses exist, people may be led to discredit part or all of that evidence. For example, consider the stimulus (6:4,1) in the hit-and-run scenario. Here, two hypotheses are supported at a higher level than the third. Indeed, the weak support for the third hypothesis suggests that it could be ignored since one could easily imagine that one witness in eleven was mistaken (cf. Kahneman & Tversky, 1982). Thus, if part or even all of the negative evidence is discredited by such an "explanation," the net strength of position f will be increased.

The above discussion indicates that the diffusion of negative evidence involves two conflicting forces: first, the number of different categories of evidence suggests a highly variable process that increases uncertainty in all positions (negative diffusion); second, the distribution of evidence can suggest that certain categories of negative evidence be discredited such that the f position is strengthened (positive diffusion). A further factor, however, can reduce both effects; viz., the similarity of the categories/ hypotheses making up the negative evidence. That is, if the negative categories are seen as highly similar, the splitting of such evidence into multiple categories will be seen as artificial and have no effect.

We now consider a model of the diffusion effect that extends our earlier work and incorporates both the conflicting and moderating forces discussed above. Let

$S_n(f:c_1,c_2,\ldots, c_j,\ldots, c_J)$ = net strength of evidence in favor of the f position, where:

c_j = amount of negative evidence in the j^{th} alternative hypothesis/category such that,

$$c = \sum_{j}^{J} c_j$$

We hypothesize that the evaluation of $S_n(f:c_1,c_2,\ldots)$ is accomplished by a multi-stage or cascaded anchoring and adjustment process in which one first anchors on $S_n(f:c)$ and then adjusts for the number of independent hypotheses/categories that comprise the negative evidence (cf. Schum, 1980). In order to model the adjustment process, denote c^* as the <u>effective</u> amount of negative evidence and assume that,

$$c^* = c + (1-\Theta)(k-d), \quad \text{where} \tag{11}$$

Θ = parameter representing the similarity of alternative categories of the negative evidence ($0 \leq \Theta \leq 1$ and $\Theta = 1$ implies maximum similarity)

k = force toward negative diffusion = increasing function of J, the number of categories of negative evidence

d = amount of negative evidence discredited.

Given equation (11), a model for the evalution of $S_n(f:c_1,c_2, \ldots, c_J)$ can be written as,

$$S_n(f:c_1,c_2, \ldots,c_J) = S_n(f:c^*) = \frac{f}{f + c^*} - \frac{a_2}{f + c^*} \tag{12}$$

Substituting (11) into (12) and combining terms yields,[5]

$$S_n(f:c_1,c_2, \ldots,c_J) = \frac{f - a_2}{n + (1-\Theta)(k-d)} \tag{13}$$

Equation (13) represents the multi-stage or cascaded anchoring and adjustment process described above. We can now define diffusion effects as being a non-zero difference between $S_n(f:c)$ and $S_n(f:c_1,c_2, \ldots,c_J)$. Specifically, using (2) and (13), let,

$\Delta = S_n(f:c) - S_n(f:c_1,c_2, \ldots,c_J)$, which can be shown to equal,

$$\Delta = \left(p - \frac{a_2}{n}\right)\left[1 - \frac{n}{n + (1-\Theta)(k-d)}\right] \tag{14}$$

Equation (14) is consistent with our earlier discussion in that no diffusion

effects would be predicted if either $\theta = 1$ (i.e., maximum similarity of the categories of negative evidence), or $k = d$ (i.e., the opposing forces cancel each other). However, note that equation (14) also indicates that as n increases, $|\Delta|$ goes to zero. Hence diffusion effects would not be expected for large n. However, when n is small, p cannot be too high since there would be little c to diffuse. On the other hand, if p is low, the first term in (14) is reduced, thereby lowering $|\Delta|$. Thus, diffusion effects are most likely when n is small and p is in the middle range.

Experimental evidence

Since the origin of diffusion effects lies in imagining how a certain pattern of results could occur in a manner different from standard outcomes (non-diffused stimuli), any effects are likely to be highly individualized and difficult to demonstrate. We therefore used four different ways to test for the effect: (a) a between-subjects design using two different scenarios; (b) a within-subjects design using a single scenario; (c) a within-subjects ranking task of standard and diffused stimuli; and, (d) a within-subjects experiment investigating the effect of similarity (θ) on diffusion. We now consider each experiment in turn.

In study (a), subjects were asked to evaluate a scenario that was presented in either a diffused or non-diffused version. The dependent variable in all cases was a judgment of the likelihood that a particular position was true. The first scenario was as follows:

> Police were searching for clues to identify a bank robber's two accomplices. The 8 witnesses to the crime all testified to the fact that the accomplices had spoken to each other in a foreign language. However, there was disagreement concerning the language spoken. Of the witnesses, 4 said it was German, whereas 4 said that it was Italian.

How likely do you think that the language was German?

In the diffused version, the evidence was given as, "Of the witnesses, 4 said it was German, whereas 2 said French, 1 said Spanish, and 1 thought it was Italian." The number of subjects in each group was 63 and 50 for the non-diffused and diffused stimulus, respectively. The mean responses were .44 in the non-diffused group and .50 in the diffused group ($t = 2.03$, $p < .04$). Therefore, positive diffusion occurred--most likely by ignoring one or more of the single discrepant witnesses.

The second scenario was as follows:

> Seven experts in classical music were invited by a local FM radio station to demonstrate their skill in identifying composers. After hearing one piece, 3 experts said it had been written by Bach, whereas 4 stated Beethoven.

How likely do you think the piece was written by Bach?

The diffused version said, "3 experts said it had been written by Bach, whereas 2 stated Beethoven, and 2 thought it was Gershwin." Twenty-nine subjects received the non-diffused version and 35 got the diffused stimulus. The mean responses were .42 vs. .35, showing negative diffusion effects ($t = 2.11$, $p < .03$). That is, the rather sizeable difference in the critics' opinions seemed to induce a diagnosis that they were all less expert.

We now consider experiments (b) and (c). In the experimental material used in Experiment 1 (the hit-and-run scenario), 8 stimuli (plus 4 repeats) were included in which negative evidence was diffused into two or three categories (various car colors reported by different numbers of witnesses). The stimuli were chosen so that they had the same p and n as stimuli that were not diffused. For example, with $p = 0$, $n = 2$, the diffused stimulus was (0:1,1). Similarly, the other stimuli were: (2:2,2) for $p = .33$, $n = 6$; (6:6,6) for $p = .33$, $n = 18$; (2:1,1,1) for $p = .4$, $n = 5$; (4:3,3) for $p = .4$, $n = 10$; (4:2,2) for $p = .5$, $n = 8$; (6:1,1,1) for $p = .67$, $n = 9$; (8:2,2) for p

= .67, n = 12. As with the non-diffused judgments, test-retest reliability was high (average r within subjects = .92; r across all subjects and responses = .84). Since each of the 32 subjects judged both the diffused and non-diffused stimuli, we compared the mean differences in net strength between the two types of evidence for each of the 8 stimuli (using a correlated t-test). Of the 8 possible effects, 2 were significant ($p < .01$ and $p < .02$) and both demonstrated a negative diffusion effects. Moreover, these effects occurred for moderate values of p (= .4) and small n(< 10), as predicted by equation (14).

Experiment (c) used a ranking task rather than asking for likelihood judgments of individual stimuli. The rationale for this was that comparative judgments were thought to focus attention on the difference between diffused and nondiffused stimuli, resulting in larger effects. The task was given to 7 of the 32 subjects who participated in Experiment 1, with the following instructions:

> Please reconsider the automobile accident scenario. Below you will find several cards indicating the colors of the car reported by various witnesses. Consider, on the basis of the evidence on each card, how likely you believe that the offending car was green. Then rank order the cards from most to least likely. You are free to state that some cards represent equally likely situations.

Subjects were asked to rank 14 stimuli of which 12 comprised 6 pairs of diffused vs. non-diffused evidence. These pairs were: (12:6) vs. (12:3,3); (8:4) vs. (8:2,2); (6:4) vs. (6:2,2); (3:2) vs. (3:1,1); (4:4) vs. (4:2,2); (2:3) vs. (2:1,1,1). The simplest way to analyze the results of this experiment is to count the number of non-ties for the 6 pairs of interest. The results of this procedure showed that for 5 subjects, diffusion effects occurred for all 6 pairs, 1 subject showed no effects, and another showed 1 effect. Furthermore, negative diffusion effects were much more likely than

positive ones (26 vs. 5). Taken together, diffusion effects occurred in 31 of the possible 42 cases (74%).

Although we have demonstrated the existence of diffusion effects, we did not predict their direction (i.e., positive or negative). To do this would require a theory that specifies the factors that comprise the opposing forces defined in our model. However, whereas this topic is important, its detailed treatment lies beyond the scope of this paper and thus we limit ourselves to some speculative comments. First, we hypothesize that the force toward negative diffusion (k) will be positively related to the number of categories of negative evidence, J. However, this may be moderated by the extent to which people believe that they are dealing with an unstable phenomenon. Specifically, we suggest that the greater one's surprise that evidence is diffused, the greater the force toward negative diffusion. However, the "surprisingness" of evidence is obviously a highly idiosyncratic matter that depends on one's experience and knowledge of a particular content area. Thus, predicting the direction of diffusion effects will depend on knowing something about people's prior beliefs concerning particular causal processes. Second, we believe that the amount of negative evidence discredited (i.e., d) will depend on the distribution of evidence across both the favored and alternative categories. As noted above, whereas several items of evidence within a single category provide some consensus for that category, this is not true of categories containing few or single items since these can be easily "undone" (i.e., discredited) through mental simulation (cf. Kahneman & Tversky, 1982).

Our model does, however, make a simple prediction concerning the non-occurrence of diffusion effects. That is, when evidence is seen as highly

similar and overlap between categories is large, we hypothesized that $\upsilon = 1$, resulting in no effects. The following experiment (d) explicitly tested this implication.

Subjects (N = 67) were asked to read the following scenario and then respond by marking a point on a 0-100 scale:

> Psychologists are investigating a means of testing the reading skills of 6 year-old children. The procedure consists of flashing a three-letter word on a screen for half a second and subsequently asking the children what they saw. Nine children were tested. For one test word, 6 children stated they had seen ROT, whereas 3 said BED.
>
> How likely do you think that the word was ROT?

In addition to the above stimulus, each subject was asked to judge two diffused versions. The first contained highly similar alternative hypotheses; specifically, ". . . 6 children stated they had seen ROT, whereas 1 said BED, 1 said BAD and 1 said BID." According to our model, $\upsilon = 1$ for this stimulus and no diffusion effects were expected. The second diffused stimulus involved dissimilar alternatives; i.e., ". . . 1 said CUP, 1 said BED and 1 said FAR." For this stimulus we predicted diffusion effects, but not their direction. In order to control for possible order effects, half the subjects were given the stimuli in the order discussed above while the other half received the stimuli in reverse order. Finally, the diffused stimuli were given some 40 minutes apart from the standard version.

The results showed no order effects so we consider the means for the three stimuli over all subjects. These were .63, .66, and .74 for the standard version, similar alternatives and dissimilar alternatives, respectively. The difference between the first two means is not significant (as expected) while the third is significantly greater than the other two (showing positive diffusion effects, $t = 5.93$, $t = 5.79$, $p < .001$). Therefore, the results conform to the model predictions and demonstrate the importance of

alternative similarity in affecting the occurrence of diffusion effects.

Discussion

The importance of the diffusion concept goes beyond the evaluation of evidence in the simple situations just considered. To illustrate, we consider two related issues that put the concept in a much broader context: (a) The diffusion effect has an interesting connection to the idea of equating groups in experimental designs through randomization. This concept is not intuitively appealing, as anyone who has tried to teach it will know. Indeed, some researchers have difficulty in accepting it. For example, consider the following from Campbell and Stanley (1963):

> Perhaps Fisher's most fundamental contribution has been the concept of achieving pre-experimental equation of groups through randomization. This concept, and with it the rejection of the concept of achieving equation through matching (as intuitively appealing and misleading as that is) has been difficult for educational researchers to accept. (p. 2)

Why is matching intuitively appealing but randomization is not? In order to assess whether some experimental treatment causes some response, one must rule out alternative explanations. Note that this is entirely consistent with our view of evidence as a net effect; i.e., the net strength of evidence for a hypothesis is the strength of positive evidence less the strength of competing alternatives. "Matching" erroneously seems to rule out alternative hypotheses since the groups are assumed to be equal except for the experimental treatment. Randomization, on the other hand, seems to go in the opposite direction by letting groups vary in innumerable ways. However, this is the essence of the negative diffusion effect. That is, by increasing the number of factors that the groups could differ on to an immense number, one reduces to zero the net strength in favor of any <u>single</u> factor causing a difference. Therefore, by increasing the number of alternative hypotheses, diffusion occurs such that

the groups are equated. The relation between randomization and diffusion also makes clear why one must have reasonably sized samples for randomization to work--a large sample insures that enough differences are present to reduce the net strength of any single alternative to an insignificant level. Thus, the nonintuitive nature of randomization lies in the fact that more, rather than less, alternative explanations leads to the equating of experimental groups; (b) The parameter θ, which is central to the diffusion concept, plays an analogous role in the evaluation of evidence to that of "similarity" in theories of choice (e.g., Tversky, 1972). In particular, it is known that the similarity of alternatives in a choice set can affect choice probabilities so that the addition or deletion of irrelevant alternatives can nevertheless have large effects. In an analogous way, the θ parameter, which we interpreted as reflecting the similarity of alternative explanations, affects the evaluation of evidentiary strength. While a complete analysis of the relation between similarity in choice and similarity in inference is beyond the scope of this article, the fact that there seems to be a relation is a positive indication that inference and choice may rest on common principles (cf. Einhorn & Hogarth, 1981). Clearly, much work is needed to clarify the links between inferential judgments and choice, but "similarity" seems a promising place to start.

(4) The missing evidence effect

The diffusion effect illustrates the point that people do not restrict their inferential process to the available data but are quick to seize upon any cues that can illuminate the task (cf. Hammond, 1972). One such cue is the structure of negative evidence. Another is missing data, which can take

the form of indirect evidence. For example, consider a person who refuses to testify in a trial of a reputed gang member, or a referee who doesn't respond to a request for an evaluation of someone being considered for promotion. Although these cases strictly involve a lack of evidence, they can be interpreted as suggesting that the gang member is guilty, and the person is undeserving of promotion, respectively. The reason is that in both cases the non-response can be seen as diagnostic of underlying states from which nonresponses are highly likely. Therefore, guilt and lack of ability are likely states to have caused (or generated) the nonresponses, thereby changing missing evidence into negative information.

The missing evidence effect can be conceptualized as resulting from a multi-stage or cascaded anchoring and adjustment process where one first anchors on $S_n(f:c)$ and then adjusts for the missing evidence which we denote by m. Specifically, since the missing evidence is presumed to be diagnostic, it relates to either the f or the c position. Thus, if m supports the f position, define

$$f' = f + m \tag{15}$$

If m supports the negative side, define,

$$c' = c + m \tag{16}$$

Given (15), the new evidence functions, $S_n(f':c)$ and $S_n'(f':c)$, for $p > p_c$ and $p < p_c$, respectively, can be written as,

$$S_n(f':c) = \frac{f' - a_2}{f' + c} = \frac{f + m - a_2}{n + m} \tag{17a}$$

and

$$S'_n(f':c) = \frac{f' + a_2}{f' + c} = \frac{f + m + a_2}{n + m} \qquad (17b)$$

Given (16), the analogous evidence functions for $S_n(f:c')$ and $S'_n(f:c')$ are

$$S_n(f:c') = \frac{f - a_2}{f + c'} = \frac{f - a_2}{n + m} \qquad (18a)$$

and

$$S'_n(f:c') = \frac{f + a_2}{f + c'} = \frac{f + a_2}{n + m} \qquad (18b)$$

Experimental evidence

The notion that missing evidence can be diagnostic was tested in both within and between subjects experiments. In the within-subjects design, 67 subjects who participated in an earlier experiment were given two versions of the following scenario:

> A bank is checking the credentials of a job applicant. As is usual for the type of position involved, requests have been made for 4 letters of recommendation. Of the 4 requests for recommendation, 3 are positive and 1 is negative.
>
> How likely do you think this is a "positive" candidate?

In the missing evidence version of the above, it was stated that of 7 letters requested, 3 were positive, 1 was negative, and 3 people did not respond. Furthermore, half the subjects received the two scenarios in one order and half received the other order. For this scenario we hypothesized that the effect of the missing evidence would be to decrease the likelihood that the candidate was "positive." That is, the fact that referees failed to respond could be attributed to their not wanting to write a negative letter.

Since order of presentation made no difference in the results, we simply consider the mean responses to the two versions of the scenario. These means are .63 for the standard version and .43 for the missing evidence stimulus ($p < .001$). Clearly, the addition of 3 nonresponses was interpreted as negative evidence since the mean not only went down, but decreased to below .50. That is, the missing evidence made the stimulus look more like (3:4) than (3:1). Indeed, given Equation (18a), we can estimate that $\underset{\sim}{m} = 3$.

The between-subjects design involved our previous hit-and-run scenario as the standard (N = 32) against two versions employing missing evidence (N = 67). These latter versions were as follows:

> An automobile accident occurred at a street corner in downtown Chicago. The car that caused the accident did not stop but sped away from the scene. Although about 10 persons witnessed the accident, police were only able to trace and interview 5. Of these 5 witnesses, 3 reported that the color of the offending car was green, whereas 2 reported it was blue.
>
> It was subsequently discovered that many of the witnesses police did not interview were personal friends of the accused driver of a green car.
>
> How likely do you think that the color of the offending car was green?
> [In the second variation of this scenario, it was stated that there were about 50, instead of 10, witnesses to the accident, all other details being strictly the same.]

For these scenarios, we hypothesized that the nature of the missing evidence would increase the strength of the green car hypothesis in that subjects would infer that witnesses had reasons to avoid being questioned by the police.

The results showed that, whereas the mean response to the standard version was .54, the means for the missing evidence versions were .66 and .65, respectively ($p < .001$).

In summary, we have demonstrated empirically that missing evidence can affect estimates of evidentiary strength. In closing this section we note a

classic, albeit fictitious example of the diagnostic impact of missing evidence drawn from Conan Doyle's, Silver Blaze:

> "Is there any other point to which you wish to draw my attention?"
>
> "To the curious incident of the dog in the night-time."
>
> "The dog did nothing in the night-time."
>
> "That was the curious incident," remarked Sherlock Holmes.

General Discussion

In assessing the diagnostic impact of evidence, people are sensitive to the context in which that evidence occurs (cf. Einhorn & Hogarth, 1981; Payne, in press). Indeed, we have argued that the strength of evidence can only be understood in relation to some background or field; thus our emphasis on the net strength of evidence. Moreover, our position can be characterized as basically Gestalt in orientation, but without the attendant vagueness that sometimes accompanies that label. Such vagueness stems from the difficulty of specifying the complex relations that can exist between figure and ground. However, it is only by developing models of these complex relations that they become accessible to scrutiny and empirical test. We believe that our models for evaluating evidence meet the criteria of specificity and falsifiability and that the principles incorporated therein are of importance for understanding cognition in general and inference in particular. We now demonstrate this by discussing three aspects of our theoretical formulation that seem particularly germane for understanding how inferential judgments are made: (1) the importance of imagination in constructing and testing hypotheses; (2) the prevalence of anchoring and adjustment processes in judgment; (3) the role of attention in affecting the evaluation of evidence. Following this, we conclude by discussing some normative implications of the present work.

(1) The role of imagination in inference

The basic model for evaluating the net strength of evidence (as expressed in equation (4)), proposes a psychological process that can be likened to a "subjective sensitivity analysis." That is, the objective evidence at hand provides an anchoring point while the imagining of a different result provides a test of the strength of that evidence. Therefore, the psychological mechanism in determining net strength involves comparing evidence against an imagined contrast or background scenario and then adjusting for the result of the comparison. This is true regardless of whether the imagined contrast case is worse ($p > p_c$) or better ($p \leq p_c$). Moreover, the model assumes that evidentiary strength will be adjusted according to the amount of evidence at hand; the largest adjustments occuring at small n and decreasing as n increases. This is reasonable since it is easier to undo evidence by counter-factual reasoning and the creation of "possible worlds" when that evidence is meager (cf. Kahneman & Tversky, 1982). However, as evidence increases, it is more difficult to imagine how the results could have been otherwise. In fact, one can consider our model as reflecting a process whereby people average "what is," with "what might have been." As evidence for the former gets larger, the imaginability of the latter decreases.

While our aggregate model captures the use of imagination in inference, the importance of individual differences should not be overlooked. For example, consider the a_2 parameter, which is the relative weight given to the imagined contrast case. As shown earlier (Table 3), there is considerable variability in the estimated values of a_2. The causes of this variability remain unknown. However, we can speculate that since a_2 represents the degree to which imagination influences judgment, its variability reflects individual predispositions to engage in counter-factual reasoning and thought

trials. Whether one could consider such a predisposition a stable "trait" or simply a reflection of content knowledge in a particular situation, is an interesting question for future research. Second, there is also variability regarding p_c. We view this parameter as reflecting both individual and content differences since it indicates when evidence shifts from being probable ($p > p_c$) to improbable ($p < p_c$). That is, at some level of p, the imagined contrast case shifts from a worse to a better case. While we have no definitive explanation for the variability of p_c, we speculate that the content of various scenarios and people's varying knowledge/experience of that content, influence where the line between probable and improbable evidence is drawn. However, much work needs to be done to justify these speculations.

Although the form of the evidence function implies that greater amounts of information reduce the effects of imagination, the constructive nature of imagination can lead to important exceptions. For example, assume that an acquaintance has been accused of a minor crime. After listening to his rebuttal, you are quickly convinced that he is innocent. However, you are surprised to note the vigor with which he continues to proclaim his innocence. Indeed, after a while the very extent of his statements raises suspicions and you begin to wonder whether he is not guilty after all ("Methinks he protesteth too much"). Similar suspicions have been raised in science when data look too perfect; see, for example, Kamin (1974) on Burt's twin data, and Bishop, Fienberg, and Holland (1975) concerning the outcomes of Mendel's pea experiments. These examples illustrate that the structure of outcomes can suggest *new* hypotheses such that the diagnosis contradicts the surface meaning of the evidence. As is the case in the diffusion and missing evidence effects (see below), people seem to ask themselves the following question when confronted

by a particular pattern of outcomes, "What kind of causal process could produce these results?" If a particular pattern is discrepant with the process that supposedly produced it, this can be a cue that something is amiss. Consider, for example, the diagnosis that someone was "framed" for a crime. Such a diagnosis rests on the belief that the evidence is too consistent and obvious. Similarly, the cases of Burt and Mendel are suspicious because results having so little error seem unlikely to come from research that generally produces more error. Therefore, when discrepancies are large, one can construct new hypotheses to make the data mostly (Einhorn, 1976).

The constructive nature of imagination is also at the heart of both the diffusion and missing evidence effects. In the former, the multiple categories of negative evidence "cue" people to the possibility that the generating process is itself highly uncertain (thereby resulting in negative diffusion), or, some categories are discredited (leading to positive diffusion). In either case, the diffusion effect is particularly sensitive to one's prior knowledge and experience with the substantive content of evidence since the ability to imagine alternative scenarios will depend on such knowledge/experience. Therefore, large individual differences in evaluating diffused evidence are likely to be characteristic of this effect. Furthermore, since diffusion effects result from the exertion of mental effort to construct new explanations for a pattern of outcomes, many people may not engage in such activity because the investment in imagination is too high or they lack "imagination-capital." In either case, one would expect little or no diffusion and indeed, not all of our subjects show the effect. Of course, it could be argued that laboratory situations are not the most conducive settings for people to be engaged in imaginative thinking. If nothing else, the amount

of time required to do diagnosis is seriously limited, thereby working against any effects (cf. Hammond, Note 2). For example, recall our scenario involving music critics trying to identify the composer of a particular piece of music. We found that when the standard scenario was diffused from (3 Bach: 4 Beethoven) to (3 Bach: 2 Beethoven, 2 Gershwin), negative diffusion occurred. However, there are several alternative ways to interpret the diffused pattern of results. The most immediate is that the critics are incompetent and their opinions should be discounted. But, another interpretation is possible--viz., that both Beethoven and Gershwin wrote pieces in the style of Bach and at least some of the critics were aware of these obscure pieces. While such a scenario would also lead to a negative diffusion effect, the new hypothesis directly contradicts the assumption that the critics are incompetent. Furthermore, since the construction of this scenario is highly imaginative, we doubt that it could be accomplished quickly enough in most laboratory settings.

The missing evidence effect also depends on the use of imagination. Indeed, the effect is predicated on the same principle underlying diffusion effects; namely, missing evidence cues people to ask, what process could produce this pattern of outcomes? The ability to imagine scenarios to answer this question is thus at the heart of converting missing information into diagnostic evidence. Furthermore, it is interesting to note the relationship between this effect and Gestalt notions of "closure" and "good continuation." Recall that in perception, figures that have gaps or missing parts are often seen as whole; i.e., the gaps or missing parts have somehow been filled in. In the same way, one can consider that diagnosis provides cognitive closure by providing a coherent scenario whereby missing information is no longer missing (in the sense of being absent), but fits in and completes the constructed hypothesis.

Imagination, uncertainty, and the search for disconfirming evidence

In stressing the role of imagination in diagnostic inference, we briefly alluded to the cost of investing in imagination. Clearly, mental effort is one such cost (cf. Shugan, 1980). However, another is increased uncertainty due to the use of counter-factual reasoning and the construction of alternative scenarios. That is, by imagining the world as it might have been, uncertainty in the causes of "what is," goes up. This increased uncertainty is captured by the evidence function and the resulting subadditivity of complementary net strengths (the focus effect). However, our empirical results, as well as the entire conceptualization for our models, seem to contradict the findings having to do with people's lack of search for disconfirmatory evidence in hypothesis testing (Mynatt, et al., 1977, 1978; Wason, 1960; Einhorn & Hogarth, 1978). How can our theory, which specifically posits the use of imagined contrary evidence for assessing evidentiary strength, be reconciled with the above?

The most important difference between our tasks and those used in earlier studies is that we are concerned with causal-backward inference. In particular, the asking of a causal question seems to induce a different strategy for evaluation than asking a non-causal question. Indeed, Mackie (1974) has argued that the primitive notion of a cause involves asking oneself the following counter-factual question: Would Y have occurred if X had not? Note that this question involves a thought experiment where not-X serves as the control group. If Y occurs as often without X as it does with X, the causal significance of X is clearly in doubt. Therefore, our position is that the search for negative evidence if expanded to include the construction of alternative explanations and imagined contrast scenarios, is central to diagnostic inference. In fact, this should come as no surprise to diagnos-

ticians since they are continually involved in constructing and ruling out alternatives on the basis of symptoms, signs, and the like. Furthermore, this raises the following issue: in our tasks, conflicting evidence involved numbers of arguments for and against different specific positions. Thus, evidence for a position was not contrasted with a diffuse alternative that consisted of not-H. The importance of this is that the search for negative evidence can also be a search for confirming evidence when considering specific alternative hypotheses. For example, a physician who orders a particular test may be seeking to confirm a particular diagnosis by disconfirming others. In such a case, the meaning of "confirmation bias" is obviously problematic. On the other hand, when a hypothesis (H) can only be contrasted with a diffuse alternative such as not-H, the search for disconfirming evidence is likely to decrease.

The above argument rests on distinguishing between a "disconfirmation model" of hypothesis testing vs. a "replacement model." The latter is consistent with the Kuhnian notion that theories in science are not discarded, despite evidence to the contrary, if they are not replaced by better alternatives (Kuhn, 1962). Similarly, consider the "straw man" strategy in arguing whereby people deliberately demonstrate one line of reasoning or position to be untenable prior to replacing it with their own. We believe that the replacement view is equally strong in everyday inference. Thus, while it may be important to show that some hypothesis is incorrect, the question remains as to what did cause the result. Unless disconfirmation leads to replacement, it can only increase uncertainty and anxiety. A useful analogy might be the following: how many people would read detective stories if the author only revealed who didn't do it? The resolution of uncertainty involves hypothesis replacement, not simply hypothesis disconfirmation. Therefore, our position

is that both the causal nature of diagnosis and the specificity of alternative hypotheses make the search for negative evidence very likely.

(2) Anchoring and adjustment strategies

The prevalence and importance of anchoring and adjustment strategies proposed by Tversky and Kahneman (1974), have recently been discussed by Lopes (Note 1). She points out that averaging models, which fit a wide variety of judgmental data, can be viewed as reflecting an underlying anchoring and adjustment process. Specifically,

> ". . . averaging results from the intrinsically serial nature of multiattribute judgment, both for tasks in which the information is actually presented serially (such as the typical Bayesian task) and for tasks in which information is presented simultaneously but processed serially (such as the typical impression formation task). In either case, averaging is hypothesized to occur because subjects adopt an adjustment strategy in which they integrate new information into 'old' composite judgments by adjusting the old composite value upward or downward as necessary to make the 'new' composite lie somewhere between the 'old' composite and the value of the new information." (Lopes, p. 5)

The relevance of the above for our modeling of the assessment of evidentiary strength is obvious. However, we should stress that in our model, the "new" evidence is only imagined, rather than real. Nevertheless, its effect is to make the net strength of evidence a weighted average of p and $1/n$.

A second issue that greatly increases the prevalence and importance of anchoring and adjustment processes concerns the serial processing of information that is presented simultaneously. While it has been argued that anchoring and adjusting is the basic way of making judgments in a continuous environment (Hogarth, 1981), the fact that judgments can be made serially means that complex evidence is likely to be assessed via a series of anchorings and adjustments. Indeed, our modeling of both the diffusion and missing evidence effects involves a multi-stage or cascaded inference

process. This seems to be a useful and efficient method for assessing the strength of complex evidence since the task is effectively decomposed into smaller more manageable units. Furthermore, since the outcomes of prior anchoring and adjustments incorporates earlier information, it allows one to keep a "running total" of evidentiary strength with minimal demands on memory. While the diffusion and missing evidence stimuli are not of great complexity, we believe that when evidence comes from multiple sources (of varying credibility), and multiple time periods, cascaded anchoring and adjustment processes will be of even greater importance. The challenge for the future will be to delineate how and why certain anchors are chosen and the exact form of different adjustment processes.

(3) Attention and evidence

While a complete discussion of the importance of attention in defining evidence is beyond the scope of this paper, there are two specific issues that can be addressed: (1) the implications of attentional shifts when people use anchoring and adjustment strategies; and (2) the role of diagnosis in directing attention to new evidence. Regarding (1), consider our results on the focus effect and recall how the simple rephrasing of questions led to anchoring on either p or (1 - p). Moreover, the result of the shift in the anchor was that the subsequent adjustments led to the subadditivity of complementary net strengths. Lopes (Note 1) has similarly found that a simple change in the order in which sample information is presented can affect overall judgments by changing the anchor. For example, consider having to judge whether samples come from an urn containing predominantly red or blue balls (70/30 in both cases). You first draw a sample of 8 that shows (5R, 3B). Thereafter, you draw another sample of 8 with the result (7R:1B). After

each sample, you are asked how likely it is that you have drawn from the predominantly red urn. When the sample evidence is in the order given here, people seem to anchor on the first sample (5:3) and then adjust up for the second (stronger) sample. However, when the order of the samples is reversed, people anchor on (7:1) and adjust down for the weaker, second sample. This effect cannot be accounted for by assuming that people are using a Bayesian procedure (which treats the two situations as equal), but it does follow from an anchoring and adjustment process in which the anchor is weighted more heavily than the adjustment ($a_1 > a_2$).

The second issue concerns the interdependence of attention and diagnosis. We have already discussed how attention to a particular pattern of outcomes can lead to both diffusion and missing evidence effects. However, diagnosis can also direct attention to new sources and types of evidence, thereby directing information search. One might consider this as akin to a "top-down" information search strategy rather than the "bottom-up" type discussed in the formation of hypotheses. Moreover, the fact that both types are interacting in the formation of judgment illustrates our earlier point that inference is continually shifting between backward and forward modes. While we have stressed bottom-up information acquisition, an important top down strategy concerns being made aware of, and then searching for, unequivocal evidence; i.e., evidence that will decide the issue definitively (sometimes called "the smoking gun"). Therefore, given a tentative hypothesis, is there unequivocal evidence, and if so, where does one find it? For example, consider the response of certain members of the House of Representatives regarding Nixon's guilt in the Watergate cover up. These members argued that if Nixon was guilty, there must be unequivocal evidence on one of the tapes. Furthermore, since such evidence was not found (yet), they

would not vote for impeachment. Note that from our perspective, the missing equivocal evidence was seen as supporting innocence (or decreasing the likelihood of guilt). Indeed, if f is considered the "guilty" hypothesis, then c' = c + m, where m is the missing unequivocal evidence thought to support innocence. For others, however, the missing unequivocal evidence was seen as supporting guilt (e.g., he must have destroyed the tape because of its incriminating nature; f' = f + m). In either case, the diagnosis focused attention on the importance of this evidence, thus resulting in its search.

Normative implications

We began this paper by drawing attention to the different metaphors that can be used to study inference and in particular, the predominance of the statistical model of probability theory. However, it is important to recognize that probability theory itself often rests on an analogy--an urn from which balls are drawn at random. Various characteristics of the urn model, however, are quite different from real-world inference tasks. In particular, urns have clearly defined elements within well defined boundaries; sampling is clearly specified and the roles of luck and skill in determining outcomes can be easily distinguished. To emphasize the distinctions, consider how negative outcomes can increase one's belief in a hypothesis.

Imagine a basketball coach who, when asked about his team's chances of winning against a little known opponent, answers .50. However, after his team loses (negative datum), he says that he is much more certain that his team will win the rematch. Note that his increased confidence of winning is difficult to reconcile with a formal statistical analysis--i.e., the prior odds are even, the datum is negative yet the posterior odds increase. While such a change may be normatively questionable (cf. Ross & Lepper, 1980), one

can offer the following defense: when the coach says the odds of winning are even, he really means that he knows nothing about the other team and is therefore ignorant with respect to their kind of play. However, watching the game not only provides information in the form of an outcome, but also provides evidence about the process, i.e., the way the opposing team plays. Thus, the coach is able to make a diagnosis by watching the game and to imagine winning strategies for the rematch. Note how this process differs from the urn model in that nothing can be learned from watching the actual drawing of the balls from the urn (cf. Lopes, in press). Indeed, to model the coach's beliefs by an urn, one would have to prevent him from watching the game (i.e., only tell him the result).

The above example highlights the fact that inference involves the constant interplay between backward and forward processes. However, errors can result when the two processes are not kept conceptually distinct. One such error, aptly named the "confusion of the inverse" (Dawes, Note 3), concerns the failure to distinguish predictive from diagnostic probabilities. Eddy (1982) has documented several instances of this failing in the medical literature. For example, in interpreting mammography tests, physicians commonly confuse retrospective and predictive accuracy such that p(positive mammography|breast cancer) is incorrectly used as p(breast cancer|positive mammography). Since the base rates of breast cancer and positive mammography are quite different (the former less than the latter), the predictive accuracy of the test is considerably lower than its retrospective accuracy. The tragic consequences of confusing the inverse in this case are obvious: overprediction of cancer and many unnecessary mastectomies.

While the definition of a judgmental error seems clear in the above case, we wish to reemphasize that the manner in which one forms/constructs a

diagnosis does not lend itself to a normative comparison since explicit normative rules for diagnosis do not exist. Therefore, conditional on some diagnosis (or model), some predictions are better than others (cf. Einhorn & Hogarth, 1981). However, diagnosis itself rests on intuition and imagination.

Given that imagination affects inference, what role should it play in the evaluation of evidence? Put differently, what is the normative status of the trade-off between p and n (the evidence function), the subadditivity of complementary net strengths (the focus effect), and the diffusion and missing evidence effects? In order to answer this, one has first to specify which normative model is to be used as the standard. Whereas we discussed our results in relation to a relative frequency notion of probability, it is important to note that the Bayesian approach specifically allows for the blending of observations and imagination by way of prior distributions. For example, many forms of prior distributions are consistent with the notion that p should trade-off with n. Thus, in observing a Bernoulli process, a prior distribution over the parameter $\tilde{p}$, when combined with sample evidence, would generally not imply that f/n was the best point estimator. However, if the sample size were increased, the relative weight accorded to the prior distribution would decrease such that the estimator would approach f/n as $n \to \infty$.[6] Moreover, the fact that p is approached asymptotically from below rather than from above could reflect a conservative loss function; e.g., people do not wish to overstate the strength of conflicting evidence. In a similar way, results akin to the missing evidence effect can be obtained within a formal Bayesian analysis where a prior distribution over the missing evidence is blended with the data actually observed. Indeed, such methods have been specifically suggested for handling non-responses in surveys (Ericson, 1967).

Finally, the very assessment of prior distributions often relies on the use of imagination. For example, Good (1965) and Winkler (1967) have proposed methods whereby people are asked to weight prior information by the number of hypothetical observations they can imagine having seen (the "device of imaginary results"). Similarly, other authors have specifically considered priors that are non-data-based (Zellner, 1971). Thus, the flexibilities afforded by the Bayesian model have much in common with our descriptive model.

On the other hand, the diffusion and focus effects are not well handled by either the classical or Bayesian approaches. In the former, the assessment of evidence depends on the diagnosis suggested by that evidence. This circularity, while objectionable on purely logical grounds, nevertheless emphasizes that outcomes are considered in relation to the causal processes presumed to generate them. Moreover, both positive and negative diffusion seem justified when the structure of evidence strongly suggests an explanation that makes the data consistent or coherent with the diagnosis. This raises an important issue that can only be mentioned briefly; viz., are scenarios that are coherent more likely to be true than those that aren't? We believe that coherence is a valid, albeit imperfect cue to truth but can offer no evidence for this assertion. Finally, the fact that randomization seems to work via negative diffusion is a strong argument for the "normativeness" of the effect, at least under certain conditions.

We consider the focus effect last since it is the most problematic from a normative viewpoint. Specifically, the subadditivity of complementary net strengths that results when $a_2 > 0$ could lead to a so-called "Dutch book" (i.e., a bet where one loses regardless of the outcome). While this is undesirable, we also feel that the "weight of evidence" should matter in assessing the uncertainties associated with complementary events. In this

regard we feel sympathy both for Cohen's position that the probabilities of complementary events need not sum to one and its opposite. Therefore, our own conflict on this issue remains unresolved, highlighting the difficulties in defining what is "normative" (Einhorn & Hogarth, 1981).

While we believe that imagination should play an important part in evaluating evidence, we realize that little is known about how imagination is engaged in creating and synthesizing hypotheses (cf. Bronowski, 1978).[7] Nevertheless, given its importance, descriptive research on determining how and why processes operate as they do is a necessary condition for tackling the normative questions. Indeed, the dependence of normative rules on accurate descriptive theories is succinctly captured by the statement, "ought implies can," (Goldman, 1978).

Conclusion

In considering the evaluation of evidence, we have relied on a perceptual metaphor. In particular, an analogy was drawn between the perceptual concept of figure/ground and the idea that evidence is assessed according to its net strength. However, unlike the perceptual analogy, we have argued that the ground for evaluating the strength of evidence often comes from our imagination. Thus, our main thesis has been that imagination is essential for testing and constructing diagnoses. Accordingly, we have presented both theoretical arguments and experimental data in support of a model whereby people anchor on "what is" and then adjust for "what might have been." This model accounts for the trade-off between amount of evidence (n) and probability (p) as well as the subadditivity of complementary net strengths. Moreover, we have shown how patterns of outcomes can suggest diagnoses that can lead to diffusion and missing evidence effects.

While we believe that our model provides a simple and parsimonious explanation for various aspects of the inference process, the plausibility of the theory will ultimately be assessed by its net strength. However, whatever specific alternative formulations are developed, our general approach, which relies on the use of alternatives and contrast cases for assessing evidence, suggests the following irony: at a global level, what can be the alternative to a theory based on alternatives?

Footnotes

This work was supported by a contract from the Office of Naval Research. We would like to thank Zvi Gilula, Joshua Klayman, Haim Mano, and Arnold Zellner for comments on earlier drafts of this paper. Support from the Lady Davis Foundation Trust (Hebrew University of Jerusalem) is also gratefully acknowledged.

[1]It is legitimate to ask why the imagined worse case does not involve the deletion of one f rather than its shift to one c. While it is not difficult to model such a process, we believe that the deletion of positive evidence is often perceived to be evidence for the contrary position and is thus seen as an increase in the con position. The model to be described leads to more adjustment for n than a model based on the deletion of f. However, in either case the form of the model is similar as are the implications.

[2]It should be noted that when $p > p_c$ but p is low, it is assumed that S_n cannot be less than zero. For example, if $p_c = .10$, $p = .11$, $a_2 = .3$ and $n = 2$, $S_n(= .11 - (.3/2))$ must be zero. However, when evidence is discrete, one cannot have $p = .11$ with $n = 2$. In terms of Figure 1, our assumption means that for cases where $S_n < 0$, all dashed lines must emanate from the origin.

[3]The functions in Figure 2 are approximations of the relation between S_n and p because p does not exist at all levels of n in the discrete evidence case. Therefore, the figure shows the general relation between S_n and p for a range of n, rather than for a specific value. Empirical evidence, to be presented later, is in accord with the general form of the relation presented here.

[4]The estimated coefficient, $\hat{a}_2$, in equation (10) is not technically the same as the coefficient in (9). The former is the weight that best (in a

least squares sense) accounts for the difference between p and S_n while the latter is the weight that best accounts for S_n only. However, the estimates obtained from (10) are sufficiently close to our theoretical expectations and avoid the multicollinearity problem mentioned in the text.

[5]When $p \leq p_c$, equations (13) and (14) are modified so that the sign of a_2 is changed from minus to plus. The implications derived from (13) and (14) do not change in any appreciable way.

[6]We would like to thank Arnold Zellner for pointing this out.

[7]We consider this at length in the sequel to this paper.

Reference Notes

1. Lopes, L. L. Averaging rules and adjustment processes: The role of averaging in inference. Department of Psychology, University of Wisconsin, December 1981.

2. Hammond, K. R. Unification of theory and research in judgment and decision making. University of Colorado working paper, 1982.

3. Dawes, R. M. How to use your head and statistics at the same time or at least in rapid alternation. University of Oregon working paper, undated.

References

Allport, F. H. Theories of perception and the concept of structure. New York: Wiley, 1955.

Bishop, Y. M. M., Fienberg, S. E., & Holland, P. W. Discrete multivariate analysis: Theory and practice. Cambridge, MA: The MIT Press, 1975.

Bronowski, J. The origins of knowledge and imagination. New Haven: Yale University Press, 1978.

Bruner, J. S. Going beyond the information given. In J. S. Bruner et al (Eds.), Contemporary approaches to cognition. Cambridge, MA: Harvard University Press, 1957.

Campbell, D. T., & Stanley, J. C. Experimental and quasi-experimental designs for research on teaching. In N. L. Gage (Ed.), Handbook of research on teaching. Chicago: Rand-McNally, 1963.

Cohen, L. J. The probable and the provable. Oxford: Clarendon Press, 1977.

Eddy, D. M. Probabilistic reasoning in clinical medicine: Problems and opportunities. In D. Kahneman, P. Slovic & A. Tversky (Eds.), Judgment under uncertainty: Heuristics and biases. New York: Cambridge University Press, 1982.

Einhorn, H. J. A synthesis: Accounting and behavioral science. Journal of Accounting Research, 1976, 14 (Supplement), 196-206.

Einhorn, H. J., & Hogarth, R. M. Confidence in judgment: Persistence of the illusion of validity. Psychological Review, 1978, 85, 395-416.

Einhorn, H. J., & Hogarth, R. M. Behavioral decision theory: Processes of judgment and choice. Annual Review of Psychology, 1981, 32, 53-88.

Ericson, W. A. Optimal sample design with nonresponse. Journal of the American Statistical Association, 1967, 62, 63-78.

Estes, W. K. The cognitive side of probability learning. Psychological Review, 1976, 83, 37-64.

Gettys, C. F., & Fisher, S. D. Hypothesis plausibility and hypothesis generation. Organizational Behavior and Human Performance, 1979, 24, 93-110.

Goldman, A. I. Epistemics: The regulative theory of cognition. The Journal of Philosophy, 1978, 75, 509-524.

Good, I. J. The estimation of probabilities: An essay on modern Bayesian methods. Cambridge, MA: The MIT Press, 1965.

Green, D. M., & Swets, J. A. Signal detection theory and psychophysics. New York: Wiley, 1966.

Hammond, K. R. Inductive knowing. In J. Royce & W. Rozeboom (Eds.), The psychology of knowing. New York: Gordon & Breach, 1972.

Hogarth, R. M. Beyond discrete biases: Functional and dysfunctional aspects of judgmental heuristics. Psychological Bulletin, 1981, 90, 197-217.

Kahneman, D., & Tversky, A. Prospect theory: An analysis of decision under risk. Econometrica, 1979, 47, 263-291.

Kahneman, D., & Tversky, A. The simulation heuristic. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), Judgment under uncertainty: Heuristics and biases. New York: Cambridge University Press, 1982.

Kamin, L. J. The science and politics of IQ. Potomac, MD: Erlbaum, 1974.

Keynes, J. M. A treatise on probability. London: Macmillan, 1921.

Kuhn, T. S. The structure of scientific revolutions. Chicago: University of Chicago Press, 1962.

Latané, B. The psychology of social impact. American Psychologist, 1981, 36, 343-356.

Lopes, L. L. Doing the impossible: A note on induction and the experience of randomness. Journal of Experimental Psychology: Learning, Memory, and Cognition, in press.

Mackie, J. L. The cement of the universe: A study of causation. Oxford: Clarendon Press, 1974.

Mynatt, C. R., Doherty, M. E., & Tweney, R. D. Confirmation bias in a simulated research environment: An experimental study of scientific inference. Quarterly Journal of Experimental Psychology, 1977, 29, 85-95.

Mynatt, C. R., Doherty, M. E., & Tweney, R. D. Consequences of confirmation and disconfirmation in a simulated research environment. Quarterly Journal of Experimental Psychology, 1978, 30, 395-406.

Payne, J. W. Contingent decision behavior: A review and discussion of issues. Psychological Bulletin, in press.

Peterson, C. R., & Beach, L. R. Man as an intuitive statistician. Psychological Bulletin, 1967, 68, 29-46.

Ross, L., & Lepper, M. R. The perseverance of beliefs: Empirical and normative considerations. In R. A. Shweder (Ed.), Fallible judgment in behavioral research: New directions for methodology of social and behavioral science, 1980, 4, 17-36.

Schum, D. A. Current developments in research on cascaded inference processes. In T. S. Wallsten (Ed.), Cognitive processes in choice and decision behavior. Hillsdale, N. J.: Erlbaum, 1980.

Shugan, S. M. The cost of thinking. Journal of Consumer Research, 1980, 7, 99-111.

Tversky, A. Elimination by aspects: A theory of choice. Psychological Review, 1972, 79, 281-299.

Tversky, A. Features of similarity. Psychological Review, 1977, 84, 327-352.

Tversky, A., & Kahneman, D. Judgment under uncertainty: Heuristics and biases. Science, 1974, 185, 1124-1131.

Tversky, A., & Kahneman, D. The framing of decisions and the psychology of choice. Science, 1981, 211, 453-458.

Wason, P. C. On the failure to eliminate hypotheses in a conceptual task. Quarterly Journal of Experimental Psychology, 1960, 12, 129-140.

Winkler, R. L. The assessment of prior distributions in Bayesian analysis. Journal of the American Statistical Association, 1967, 62, 776-800.

Zellner, A. An introduction to Bayesian inference in econometrics. New York: Wiley, 1971.

TABLE 1

Fit of the Model for Aggregate Data

(1) n	(2) p	(3) $\bar{S}_n$	(4) $\hat{S}_n$
2	1	.85	.87
5	1	.92	.95
12	1	.95	.98
20	1	.96	.99
3	.89	.88	.86
5	.89	.87	.88
(8)	(.89)	(.88)	(.88)
27	.89	.87	.88
3	.80	.80	.75
(5)	(.80)	(.79)	(.75)
11	.80	.79	.77
15	.80	.81	.79
20	.80	.80	.79
28	.80	.82	.79
(25)	(.80)	(.80)	(.79)
4	.75	.63	.69
3	.67	.61	.58
(3)	(.67)	(.59)	(.58)
5	.67	.62	.63
(6)	(.67)	(.63)	(.63)
9	.67	.61	.64
12	.67	.64	.65
15	.67	.65	.65
18	.67	.63	.66
24	.67	.66	.66
5	.60	.53	.55
10	.60	.58	.57
2	.50	.45	.37
8	.50	.44	.47
(8)	(.50)	(.47)	(.47)
12	.50	.47	.48
20	.50	.47	.49
5	.40	.36	.35
10	.40	.39	.37
5	.33	.31	.29
(6)	(.33)	(.29)	(.29)
9	.33	.27	.30
18	.33	.29	.32
4	.25	.20	.19
5	.20	.21	.25
10	.20	.19	.20
(10)	(.20)	(.18)	(.23)
9	.11	.12	.14
19	.11	.13	.12
2	0	.16	.13
5	0	.07	.04
12	0	.06	.02
20	0	.04	.01

Note: Numbers in parentheses are for the repeat judgments.

TABLE 2

Focus Effects for the Aggregate Data

p	(1 - p)	n	Actual $\bar{S}_n(f:c) + \bar{S}_n(c:f)$	Predicted $\hat{S}_n(f:c) + \hat{S}_n(c:f)$
1	0	2	1.01	1.00
1	0	6	.99	1.00
1	0	12	1.01	1.00
1	0	20	.99	1.00
.89	.11	9	1.00	1.00
.89	.11	18	1.00	1.00
(.89)	(.11)	(18)	(.98)	(1.00)
.80	.20	5	1.01	1.00
(.80)	(.20)	(5)	.94	1.00
.80	.20	10	.98	1.00
.75	.25	4	.83	.88
.67	.33	6	.92	.92
(.67)	(.33)	(6)	(.92)	(.92)
.67	.33	9	.88	.94
.67	.33	18	.92	.98
.60	.40	5	.89	.90
.60	.40	10	.97	.94
.50	.50	2	.90	.74
.50	.50	8	.88	.94
(.50)	(.50)	(8)	(.93)	(.94)
.50	.50	12	.95	.96
.50	.50	20	.93	.98

Note: Numbers in parentheses are for the repeat judgments.

TABLE 3

Estimates of $\hat{a}_2$ in Regressions of $(p - S_n)$ on $(\frac{1}{n})$

S_s	$\hat{a}_2$	t	p_c
1	.52	4.9	0
2	.11	3.9	0
3	.32	5.5	0
4	.15	2.3	.11
5	.22	4.9	.33
6	.50	5.6	.33
7	.29	10.0	0
8	.52	9.6	0
9	.28	3.8	0
10	.46	6.3	.50
11	.09	1.6	0
12	.20	2.9	0
13	.10	.9	0
14	-.06	-1.0	0
15	.14	4.2	0
16	.15	3.1	0
17	.09	1.5	0
18	.02	2.3	0
19	.05	1.6	0
20	.08	1.0	.40
21	.06	4.1	0
22	.04	.7	0
23	.27	9.9	0
24	.24	7.3	0
25	.73	8.1	.11
26	.33	4.4	0
27	.03	1.2	0
28	.33	7.4	0
29	.31	4.0	0
30	.31	4.2	0

Note: $t = 2.0$, $p < .05$; $t = 2.4$, $p < .01$, $t = 3.5$, $p < .001$

TABLE 4

Focus Effects for Three Subjects

			S_{27} $\hat{a}_2 = 0,\ p_c = 0$		S_{24} $\hat{a}_2 = .24,\ p_c = 0$		S_{25} $\hat{a}_2 = .73,\ p_c = .11$	
p	(1-p)	n	Actual	Predicted	Actual	Predicted	Actual	Predicted
1	0	2	.99	1.00	1.00	1.00	.70	1.00
1	0	6	.99	1.00	.90	1.00	1.01	1.00
1	0	12	1.04	1.00	.99	1.00	.95	1.00
1	0	20	.92	1.00	.98	1.00	.95	1.00
.89	.11	9	1.01	1.00	.99	.94	.89	1.00
.89	.11	18	.97	1.00	.99	.98	.83	1.00
.80	.20	5	1.00	1.00	.89	.90	.71	.72
.80	.20	10	1.00	1.00	.87	.96	.74	.86
.75	.25	4	.92	1.00	.89	.88	.59	.64
.67	.33	6	1.01	1.00	.99	.92	.59	.76
.67	.33	9	1.01	1.00	.99	.94	.60	.84
.67	.33	18	.99	1.00	.99	.98	.67	.92
.60	.40	5	1.00	1.00	.99	.90	.71	.72
.60	.40	10	.90	1.00	.79	.95	.68	.86
.50	.50	2	1.00	1.00	.88	.76	.81	.27
.50	.50	8	1.00	1.00	.89	.94	.70	.82
.50	.50	12	1.00	1.00	.90	.96	.80	.88
.50	.50	20	1.00	1.00	.90	.98	.76	.92

TABLE 5

Analysis of Aggregate Data for Alternative Scenarios

Scenario	(1) $\hat{a}_2$	(t)	(2) p_c	(3) r_{p,S_n}	(4) $r_{(p-S_n),\ 1/n}$
Bank Robbery	.22	(6.34)	0	.99	.22
FM Station	.44	(9.85)	.11	.97	.63
Fire	.14	(3.62)	.25	.99	.23
Word Experiment	.21	(7.25)	.40	.99	.48

Note: $t = 2.4$, $p < .01$.

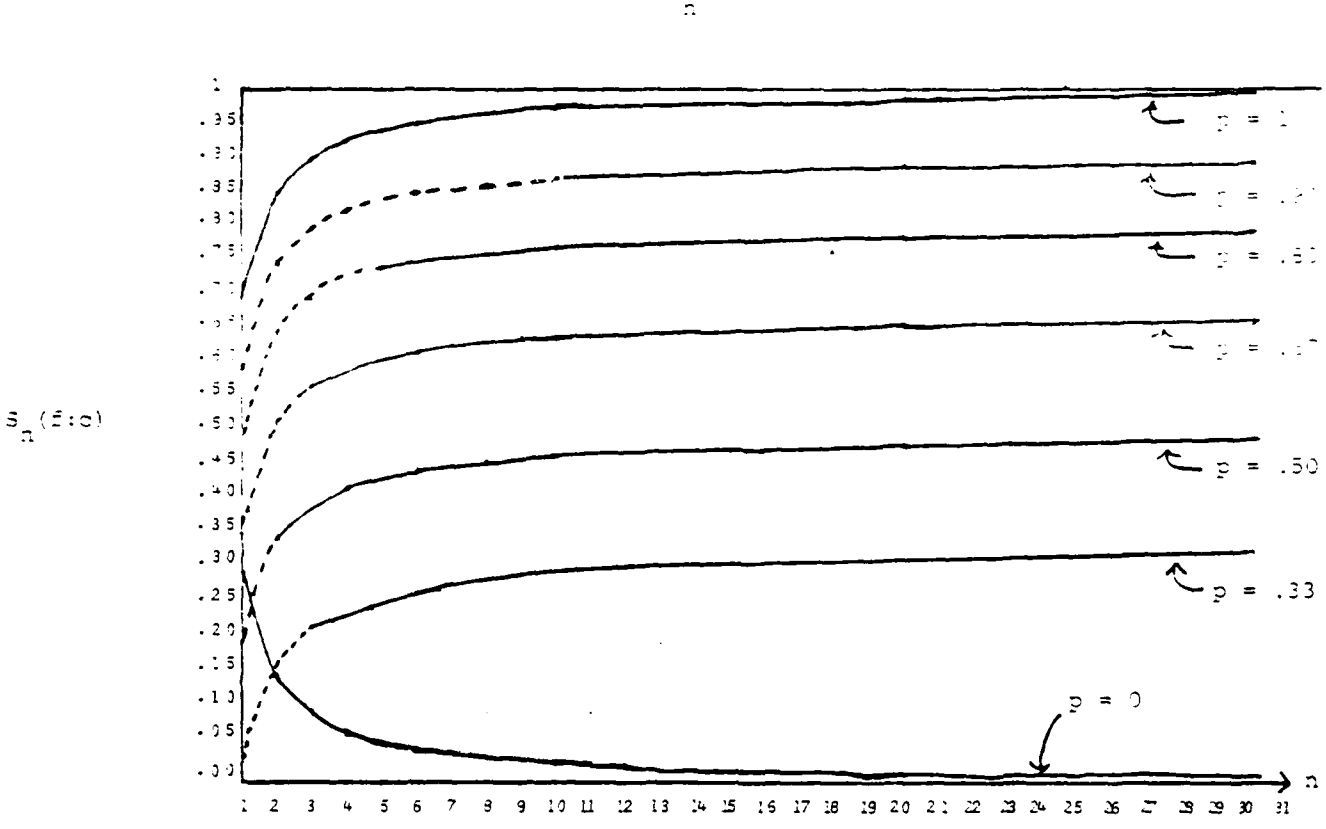


Figure 1: The Evidence Function

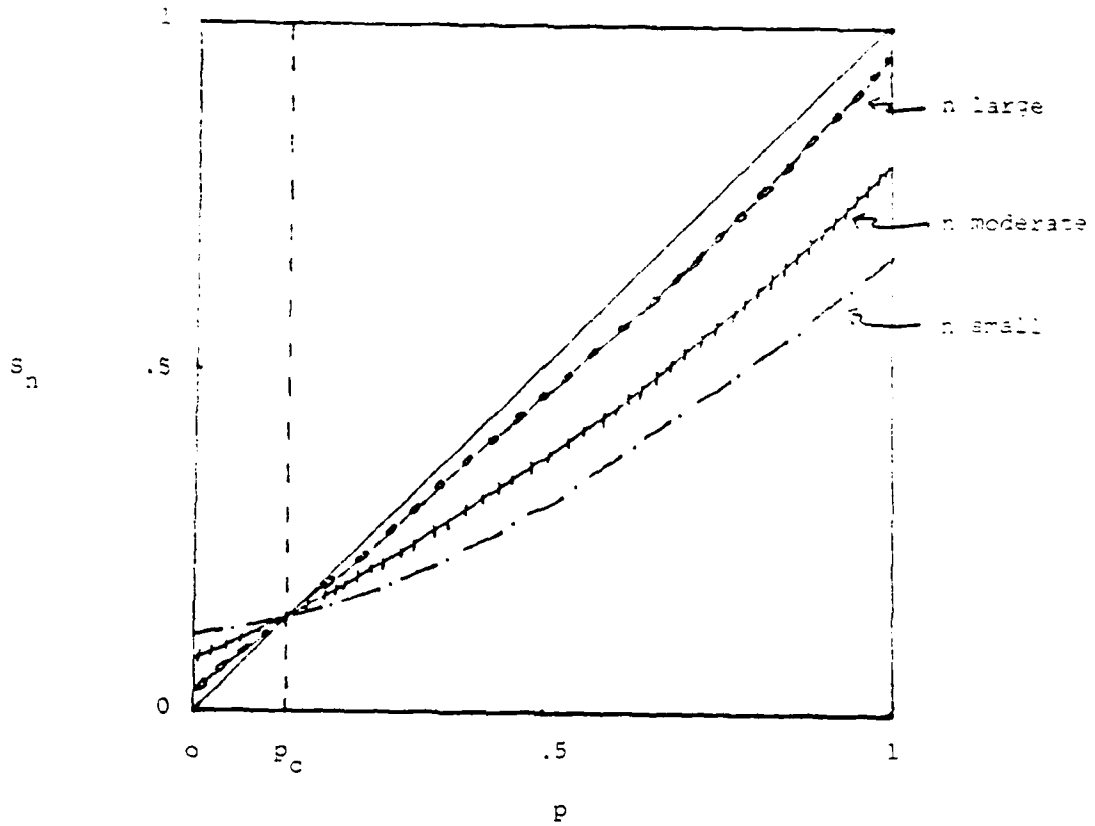
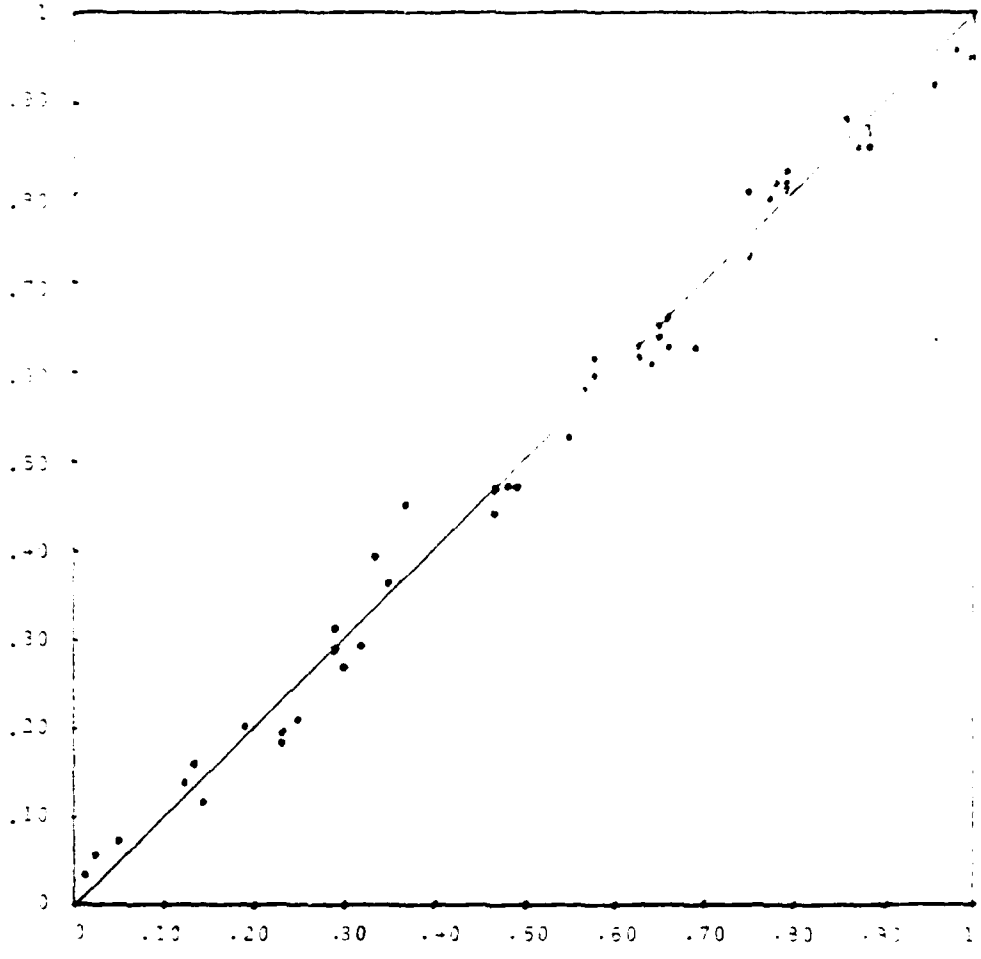


Figure 2: Net Strength as a Function of p, at varying sizes of n.



$$\hat{S}_n(f;c) = p - \beta_1(.26/n)$$

$$p_c = .20$$

Figure 3: Actual vs. Predicted Net Strength for Aggregate Data

TECHNICAL REPORTS DISTRIBUTION LIST

CDR Paul R. Chatelier
Office of the Deputy Under
Secretary of Defense
OUSDRE (E&LS)
Pentagon Room 3D129
Washington, DC 20301

Dr. Stuart Starr
Office of the Assistant Secretary
of Defense (C3I), Pentagon
Washington, DC 20301

Engineering Psychology Programs
Code 442
Office of Naval Research
800 North Quincy Street
Arlington, VA 22217 (5 cys)

Aviation & Aerospace Technology
Programs, Code 210
Office of Naval Research
800 North Quincy Street
Arlington, VA 22217

Communication & Computer Technology
Programs, Code 240
Office of Naval Research
800 North Quincy Street
Arlington, VA 22217

Tactical Development & Evaluation
Support Programs, Code 230
Office of Naval Research
800 North Quincy Street
Arlington, VA 22217

Manpower, Personnel and Training
Code 270
Office of Naval Research
800 North Quincy Street
Arlington, VA 22217

Mr. Randy Simpson
Code 411 MA
Office of Naval Research
800 North Quincy Street
Arlington, VA 22217

Statistics and Probability Program
Code 411-S&P
Office of Naval Research
800 North Quincy Street
Arlington, VA 22217

Information Sciences Division
Code 433
Office of Naval Research
800 North Quincy Street
Arlington, VA 22217

CDR K. Hull
Code 230
Office of Naval Research
800 North Quincy Street
Arlington, VA 22217

Physiology & Neuro Biology Programs
Code 441B
Office of Naval Research
800 North Quincy Street
Arlington, VA 22217

Special Assistant for Marine Corps Matters
Code 100M
Office of Naval Research
800 North Quincy Street
Arlington, VA 22217

Commanding Officer
ONR Eastern/Central Regional Office
ATTN: Dr. J. Lester
Building 114, Section D
666 Summer Street
Boston, MA 02210

Commanding Officer
ONR Western Regional Office
ATTN: Dr. E. Gloye
1030 East Green Street
Pasadena, CA 91106

Director
Naval Research Laboratory
Technical Information Division
Code 2627
Washington, DC 20375

Dr. Michael Melich
Communications Sciences Division
Code 7500
Naval Research Laboratory
Washington, DC 20375

Dr. Robert G. Smith
Office of the Chief of Naval Operations
OP987H, Personnel Logistics Plans
Washington, DC 20350

Dr. W. Mehuron
Office of the Chief of Naval
Operations, OP 987
Washington, DC 20350

Dr. Jerry C. Lamb
Combat Control Systems
Naval Underwater Systems Center
Newport, RI 02840

Human Factors Department
Code N215
Naval Training Equipment Center
Orlando, FL 32813

Dr. Alfred F. Smode
Training Analysis and Evaluation Group
Naval Training Equipment Center
Code N-00T
Orlando, FL 32813

Cdr. Norman F. Lane
Code N-7A
Naval Training Equipment Center
Orlando, FL 32813

Mr. Milon Essoglou
Naval Facilities Engineering Command
R&D Plans and Programs
Code 03T, Hoffman Building II
Alexandria, VA 22332

CDR Robert Biersner
Naval Medical R&D Command
Code 44
Naval Medical Center
Bethesda, MD 20014

Dr. Arthur Bachrach
Behavioral Science Department
Naval Medical Research Institute
Bethesda, MD 20014

CDR Thomas Berghage
Naval Health Research Center
San Diego, CA 92152

Dr. George Moeller
Human Factors Engineering Branch
Submarine Medical Research Lab
Naval Submarine Base
Groton, CT 06340

Head, Aerospace Psychology Department
Code L5, Naval Aerospace Medical
Research Laboratory
Pensacola, FL 32508

Dr. James McGrath
CINCLANT FLT
Code 04E1
Norfolk, VA 23511

Dr. Robert Blanchard
Navy Personnel Research
and Development Center
Command and Support Systems
San Diego, CA 92152

LCDR Stephen D. Harris
Human Factors Engineering Division
Naval Air Development Center
Warminster, PA 18974

Dr. Julie Hopson
Human Factors Engineering Division
Naval Air Development Center
Warminister, PA 18974

Mr. Jeffrey Grossman
Human Factors Branch, Code 3152
Naval Weapons Center
China Lake, CA 93555

Human Factors Engineering Branch
Code 1226
Pacific Missile Test Center
Point Mugu, CA 93042

Human Factors Section
Systems Engineering Test
Directorate
U.S. Naval Air Test Center
Patuxent River, MD 20670

Human Factor Engineering Branch
Naval Ship Research and Development
Center, Annapolis Division
Annapolis, MD 21402

Dr. Gary Poock
Operations Research Department
Naval Postgraduate School
Monterey, CA 93940

Dean of Research Administration
Naval Postgraduate School
Monterey, CA 93940

Mr. H. Talkington
Ocean Engineering Department
Naval Ocean Systems Center
San Diego, CA 92152

Mr. Warren Lewis
Human Engineering Branch
Code 8231
Naval Ocean Systems Center
San Diego, CA 92152

Dr. Ross L. Pepper
Naval Ocean Systems Center
Hawaii Laboratory
P. O. Box 997
Kailua, HI 96734

Dr. A. L. Slafkosky
Scientific Advisor
Commandant of the Marine Corps
Code RD-1
Washington, DC 20380

Ms. Dianne Davis
Code 3512, Building 1171
Naval Underwater Systems Center
Newport, RI 02840

Commander
Naval Air Systems Command
Human Factors Program
NAVAIR 340F
Washington, DC 20361

Mr. Phillip Andrews
Naval Sea Systems Command
NAVSEA
Washington, DC 20362

CDR W. Moroney
Code 55MP
Naval Postgraduate School
Monterey, CA 93940

Dr. Joseph Zeidner
Technical Director
U.S. Army Research Institute
5001 Eisenhower Avenue
Alexandria, VA 22333

Director, Organizations and Systems
Research Laboratory
U.S. Army Research Institute
5001 Eisenhower Avenue
Alexandria, VA 22333

Technical Director
U.S. Army Human Engineering Laboratories
Aberdeen Proving Ground, MD 21005

U.S. Air Force Office of
Scientific Research
Life Sciences Directorate, NL
Bolling Air Force Base
Washington, DC 20332

Chief, Systems Engineering Branch
Human Engineering Division
USAF AMRL/HES
Wright-Patterson AFB, OH 45433

Dr. Earl Alluisi
Chief Scientist
AFHRL/CCN
Brooks, AFB, TX 78235

Dr. Kenneth Gardner
Applied Psychology Unit
Admiralty Marine Technology Establishment
Teddington, Middlesex TW11 0LN, ENGLAND

Director, Human Factors Wing
Defence & Civil Institute of
Environmental Medicine
Post Office Box 2000
Downsview, Ontario M3M 3B9, CANADA

Dr. A. D. Baddeley, Director
Applied Psychology Unit
Medical Research Council
15 Chaucer Road
Cambridge, CB2 2EF ENGLAND

Dr. Daniel Kahneman
Department of Psychology
University of British Columbia
Vancouver, BC V6T 1W5 CANADA

Defense Technical Information Center
Cameron Station, Building 5
Alexandria, VA 22314 (12 cys)

Dr. Craig Fields, Director
System Sciences Office
Defense Advanced Research Projects Agency
1400 Wilson Boulevard
Arlington, VA 22209

Dr. Lloyd Hitchcock
Federal Aviation Administration
ACT 200
Atlantic City Airport, NJ 08405

Dr. M. Montemerlo
Human Factors & Simulation Technology
RTE-6, NASA HQS
Washington, DC 20546

Dr. Robert R. Mackie
Human Factors Research, Inc.
5775 Dawson Avenue
Goletta, CA 93017

Dr. Gary McClelland
Institute of Behavioral Sciences
University of Colorado
Boulder, CO 80309

Dr. H. McI. Parsons
Human Resources Research Office
300 North Washington Street
Alexandria, VA 22314

Dr. Jesse Orlansky
Institute for Defense Analyses
1801 North Beauregard Stret
Alexandria, VA 22311

Professor Howard Raiffa
Graduate School of Business Administration
Harvard University
Soldiers Field Road
Boston, MA 02163

Dr. T. B. Sheridan
Department of Mechanical Engineering
Massachusetts Institute of Technology
Cambridge, MA 02139

Dr. Arthur I. Siegel
Applied Psychological Services, Inc.
404 East Lancaster Street
Wayne, PA 19087

Dr. Paul Slovic
Decision Research
1201 Oak Street
Eugene, OR 97401

Dr. Harry Synder
Department of Industrial Engineering
Virginia Polytechnic Institute
and State University
Blacksburg, VA 24061

Dr. Amos Tversky
Department of Psychology
Stanford University
Stanford, CA 94305

Dr. Robert T. Hennessy
NAS - National Research Council
2101 Constitution Avenue, N.W.
Washington, DC 20418

Dr. Amos Freedy
Perceptronics, Inc.
6271 Variel Avenue
Woodland Hills, CA 91364

Dr. Robert Williges
Human Factors Laboratory
Virginia Polytechnic Institute
and State University
130 Whittemore Hall
Blacksburg, VA 24061

Dr. Alphonse Chapanis
Department of Psychology
The Johns Hopkins University
Charles and 34th Streets
Baltimore, MD 21218

Dr. Elizabeth Kruesi
General Electric Company
Information Systems Programs
1755 Jefferson Davis Highway
Arlington, VA 22202

Dr. Ward Edwards, Director
Social Science Research Institute
University of Southern California
Los Angeles, CA 90007

Dr. Charles Gettys
Department of Psychology
University of Oklahoma
455 West Lindsey
Norman, OK 73069

Dr. Kenneth Hammond
Institute of Behavioral Science
University of Colorado
Room 201
Boulder, CO 80309

Dr. James H. Howard, Jr.
Department of Psychology
Catholic University
Washington, DC 20064

Dr. William Howell
Department of Psychology
Rice University
Houston, TX 77001

Dr. Christopher Wickens
Department of Psychology
University of Illinois
Urbana, IL 61801

Mr. Edward M. Connelly
Performance Measurement Associates, Inc.
410 Pine Street, S. E., Suite 300
Vienna, VA 22180

Dr. Edward R. Jones, Chief
Human Factors Engineering
McDonnell-Douglas Astronautics Company
St. Louis Division
Box 516
St. Louis, MO 63166

Dr. Babur M. Pulat
Department of Industrial Engineering
North Carolina A&T State University
Greensboro, NC 27411

Dr. Richard W. Pew
Information Sciences Division
Bolt Beranek & Newman, Inc.
50 Moulton Street
Cambridge, MA 02238

Dr. Douglas Towne
University of Southern California
Behavioral Technology Laboratory
3716 South Hope Street
Los Angeles, CA 90007

Dr. John Payne
Graduate School of Business Administration
Duke University
Durham, NC 27706

Dr. Baruch Fischhoff
Decision Research
1201 Oak Street
Eugene, OR 97401

Dr. Andrew P. Sage
School of Engineering and Applied Science
University of Virginia
Charlottesville, VA 22901

Dr. Lola Lopes
Department of Psychology
University of Wisconsin
Madison, WI 53706

Dr. Stanley N. Roscoe
New Mexico State University
Box 5095
Las Cruces, NM 88003

Mr. Joseph G. Wohl
Alphatech, Inc.
3 New England Industrial Park
Burlington, MA 01803

Dr. Rex Brown
Decision Science Consortium
7700 Leesburg Pike, Suite 721
Falls Church, VA 22043

Dr. Wayne Zachary
Analytics, Inc.
2500 Maryland Road
Willow Grove, PA 19090

Dr. William R. Uttal
Institute for Social Research
University of Michigan
Ann Arbor, MI 48109

END

DATE
FILMED

07-82

DTIC